Universidade de Lisboa
Faculdade de Ciências

Departamento de Estatística e Investigação Operacional

Síndrome Coronário Agudo: Análise do impacto das variáveis sócio-demográficas, ambientais e clínicas na demora média entre o início da sintomatologia e o restabelecimento do fluxo

Daisy Andreína Vieira De Abreu

Dissertação de Mestrado

Mestrado em Bioestatística

2013

Universidade de Lisboa
Faculdade de Ciências

Departamento de Estatística e Investigação Operacional

Síndrome Coronário Agudo: Análise do impacto das variáveis sócio-demográficas, ambientais e clínicas na demora média entre o início da sintomatologia e o restabelecimento do fluxo

Daisy Andreína Vieira De Abreu

Dissertação orientada por:

Prof. Doutora Maria Salomé Cabral

Dissertação co-orientada por:

Mestre Fernando Ribeiro

Mestrado em Bioestatística

2013

Agradecimentos

Manifesto aqui o mais genuíno agradecimento à Professora Doutora Salomé Cabral, minha orientadora, pela sua inigualável disponibilidade e que de forma sempre muito sensata me orientou neste trabalho, contribuindo de maneira decisiva para o meu crescimento nesta área.

Quero expressar também o meu agradecimento e reconhecimento ao Professor Doutor Fernando Ribeiro, pela disponibilidade demonstrada a cada momento, tendo um papel preponderante na minha formação clínica e científica.

Aos doentes e familiares expresso a minha gratidão pela boa-vontade e paciência com que participaram neste estudo.

Agradeço aos meus amigos todo o apoio, ânimo e força que me deram durante o processo crítico de realização deste trabalho.

Por último, dirijo um agradecimento especial aos meus pais e irmão, por serem modelos de coragem, pelo apoio incondicional, pela disponibilidade para me ouvirem, pelo afecto, pelo incentivo, pela paciência e ajuda na superação de obstáculos que ao longo desta caminhada foram surgindo. A eles dedico este trabalho.

ÍNDICE GERAL

Índice de Tabelas

Índice de Figuras

Resumo

O Síndrome Coronário Agudo (SCA), é a doença com maior taxa de mortalidade e morbilidade nos países desenvolvidos, sendo a segunda causa de morte mais frequente em Portugal. O enfarte agudo do miocárdio (EAM) constitui a manifestação mais grave do SCA, e requer intervenção médica urgente para melhorar a sobrevivência e a qualidade de vida dos sobreviventes. Quanto mais precocemente for realizado o tratamento menor o tempo total de isquémia, que é definido como o tempo desde o início da sintomatologia até ao início do tratamento. Na maioria dos estudos foi demonstrado que um aumento do tempo total de isquémia estava associado a um pior prognóstico. Tendo em conta que os doentes chegam tardiamente ao tratamento, é importante reconhecer quais são os factores que condicionam o atraso no tratamento. Esta tese tem como objectivo a identificação desses factores/variáveis a partir da análise de um conjunto de dados recolhidos no Serviço de Cardiologia I do Hospital de Santa Maria. A regressão logística foi a metodologia estatística utilizada e os dados foram analisados usando o *software* R versão 2.13.

Para a obtenção do modelo de regressão logística final foram utilizadas varias técnicas de selecção de covariáveis: método de selecção de covariáveis "Hosmer e Lemeshow" e o método *stepwise.* Depois de obtido o modelo foi verificado o seu ajuste ao conjunto de dados e avaliada a sua capacidade preditiva.

O modelo final revelou seis covariáveis associadas à variável resposta, tempo total de isquémia, que foram: idade do doente, o nível de intensidade da dor, a zona de proveniência, o nível socioeconómico, as funções que se encontrava a realizar no momento de instalação do quadro, e por último o facto de o doente ter sido transferido de outro hospital.

Pode-se assim concluir que a análise do conjunto de dados através da regressão logística possibilitou a identificação das covariáveis associadas ao tempo total de isquémia. A identificação destas covariáveis permite ainda a identificação dos doentes que constituem um grupo com possibilidade de pior prognóstico, para os quais devem ser dirigidos os esforços educacionais.

Palavras-chave: Síndrome Coronário Agudo, Enfarte Agudo do Miocárdio, Tempo Total de Isquémia, Modelo de Regressão Logística.

Abstract

The Acute Coronary Syndrome (ACS), is the disease with the highest mortality and morbidity rate in developed countries and the second most frequent cause of death in Portugal. The acute myocardial infarction (AMI) is the most serious manifestation of the ACS, and requires urgent medical intervention to improve survival and quality of life of survivors. The sooner the treatment is performed the less the total time of ischemia, which is defined as the time from the onset of symptoms until treatment is achieved. In most studies it was shown that an increase in the total time of ischemia was associated with a worse prognosis. Given that most patients arrive late for treatment, it is important to understand which factors influence the delay in treatment. The main gold of this thesis is the identification of those factors. Data from the "Serviço de Cardiologia I do Hospital de Santa Maria" were analysed using the logistic regression approach, using R software, version 2.13.

To obtain the final logistic regression model, several techniques of covariates selection have been applied, such as the method of selection of covariates "Hosmer and Lemeshow" and the *stepwise* method. After the final model was obtained, the fit of the model was assessed and its predictive ability was evaluated.

The final model revealed six covariates associated with the response variable, total time of ischemia, which were: patient age, level of pain intensity, the area of origin, socioeconomic status, functions that the patient was performing at the time of installation of symptoms, and finally the fact that the patient has been transferred from another hospital.

In conclusion, the application of logistic regression to data set allowed the identification of covariates associated with the total time of ischemia, some of which can be modified to optimize the therapy. The identification of these covariates also allows the identification of patients with possibility of worse prognosis, for which should be directed educational efforts.

Keywords: Acute Coronary Syndrome, Acute Myocardial Infarction, Total Ischemia Time, Logistic Regression Model.

Lista de abreviaturas

AI - Angina instável

AHA - American Heart Association

CI - Cardiopatia Isquémica

EAM - Enfarte Agudo do Miocárdio

EAMEST - Enfarte Agudo do Miocárdio com elevação do segmento ST

EAMSEST - Enfarte Agudo do Miocárdio sem elevação do segmento ST

ECG - Electrocardiograma

HSM - Hospital de Santa Maria

IC - Intervalo de Confiança

ICP - Intervenção Coronária Percutânea

INE - Instituto Nacional de Estatística

MLGs - Modelos Lineares Generalizados

MLAG - Modelo Linear Aditivo Generalizado

OMS - Organização Mundial da Saúde

OR - *Odds ratio*

ROC - *Reciving operating curve*

SCA - Síndrome Coronário Agudo

SU - Serviço de Urgência

Capítulo 1

Introdução

A estatística é uma ciência que pode estudar mútilplas questões, nomedamente biomédicas, permitindo a análise e interpretação estatística de parâmetros fisiológicos e factos relacionados, com o objectivo de responder a questões práticas o que a torna numa importante estratégia da investigação clínica.

Dentro da área da investigação clínica a identificação dos factores que contribuem para que um determinado fenómeno ocorra, ou não, é fundamental para melhor compreender o fenómeno em causa e, ao mesmo tempo, permitir optar por estratégias que possam melhorar a prevenção e a prática clínica.

1.1 OBJECTIVOS

Este trabalho tem como objectivo identificar os factores associados ao aumento do tempo total de isquémia cardíaca em doentes com diagnóstico confirmado de Síndrome Coronário Agudo (SCA) e com restabelecimento do fluxo sanguíneo das artérias coronárias por intervenção coronária percutanêa primária (ICP), que recorreram ao Hospital de Santa Maria no período compreendido entre 01 de Janeiro de 2010 e 31 de Dezembro de 2010, através da análise retrospectiva de processos hospitalares e contacto por via telefónica de cada um dos doentes que entraram no estudo. O modelo de regressão logística foi a metodologia estatística utilizada.

Até ao presente, não existe nenhum estudo, para a população portuguesa que permita a identificação dos factores associados ao aumento do tempo total de isquémia cardíaca, o que denota a extrema importância do presente estudo, que visa melhorar o entendimento do problema, além de fornecer dados importantes para o planeamento de acções direcionadas à educação tanto dos doentes como dos profissionais de saúde.

Para facilitação da organização e apresentação do trabalho, este foi dividido em 5 capítulos.

O Capítulo 2 é dedicado a apresentar a revisão teórica da patologia em estudo, resultante da pesquisa bibliográfica, sobre os aspectos considerados mais pertinentes para introduzir a

investigação clínica, cujo tema é o SCA e seu tratamento, mais especificamente a ICP. É feita uma breve revisão da patologia, isto é, do Síndrome Coronário Agudo, a sua definição e tratamento, o impacto da doença na sociedade, assim como se demonstra a necessidade do estudo em causa.

No Capítulo 3, é feita a revisão dos principais resultados teóricos relacionados com os modelos lineares generalizados, mais especificamente os modelos de regressão logística, desde os métodos de estimação associados até à interpretação dos valores obtidos.

No Capítulo 4, é descrita a metodologia implementada no estudo para a obtenção de um modelo de regressão logística, desde a descrição do processo de adquisição dos dados, a apresentação e caracterização dos mesmos até à explicação dos diferentes métodos de selecção de covariáveis, nomeadamente o método de "Hosmer e Lemeshow" e o método *stepwise* que conduziram ao modelo de regressão logística final. Por último aferimos a qualidade do modelo obtido assim como a capacidade preditiva do mesmo. Todas as análises estatísticas foram realizadas com recurso ao *software* R, versão 2.13.0.

Finalmente no Capítulo 5 é feita a interpretação e discussão dos resultados obtidos.

Capítulo 2

Síndrome Coronário Agudo

2.1 INTRODUÇÃO

Na Europa, os dados estatísticos revelam que as doenças cardiovasculares são responsáveis por cerca de metade de todas as mortes, causando só na União Europeia (UE), mais de 2 milhões de mortes anualmente, sendo considerada a primeira causa de morte; e são, também, responsáveis por 23% da morbilidade[1]. Estas patologias cardíacas são igualmente responsáveis por cerca de 2% dos gastos em saúde na UE, quase 24 mil milhões de euros, mas este valor adquire proporções ainda mais relevantes se se considerarem outros factores, como os gastos relacionados com a perda de produtividade, ascendendo o custo total a 50 mil milhões de euros[1].

O SCA é a doença com maior taxa de mortalidade e morbilidade nos países desenvolvidos, tendo sido a segunda causa de morte mais frequente em Portugal em 2002. A Organização Mundial de Saúde (OMS) prevê que até 2030 aproximadamente 23.6 milhões de pessoas morram devido a doença cardiovascular e que em 2020 o SCA se torne a causa mais comum de morte em todo o Mundo. Em 2008 foi estimado que 7.3 mortes estavam associadas ao SCA[2-4].

Dados retirados da Instituto Nacional de Estatística, afirmam que 1.3% da população portuguesa já teve um enfarte agudo do miocárdio (EAM), com mais homens que mulheres contribuindo para esta proporção[5].

O EAM constitui a manifestação mais grave do SCA, dado que corresponde à morte das células do músculo cardíaco. O facto destas células não se regenerarem, tem como consequência uma diminuição da força do coração para bombear o sangue para as diferentes partes do corpo, condicionando insuficiência cardíaca. O EAM foi considerada a terceira causa de mortalidade em Portugal em 2004 (8,7%)[6]. Apesar dos avanços no diagnóstico e tratamento do EAM, a sua mortalidade e morbilidade permanecem muito elevadas. O EAM requer intervenção

médica urgente para melhorar a sobrevivência e a qualidade de vida dos sobreviventes, dado que “Tempo é miocárdio!”

Dos diferentes tipos de EAM que podem ocorrer, o Enfarte Agudo do Miocárdio com elevação de segmento ST (EAMEST) é uma das principais causas de morte e morbilidade no mundo[3].

O tratamento do SCA passa pela adopção precoce de uma estratégia de reperfusão, uma vez que, quanto mais precocemente for realizado o tratamento melhor será o prognóstico, apresentando-se a ICP primária como a mais vantajosa das opções terapêuticas. Infelizmente, apenas uma pequena percentagem dos pacientes com SCA realizam ICP primária dentro do tempo estipulado [7-9], pelo que se torna importante perceber o motivo deste atraso. Alguns dos factores que contribuem para este atraso já foram identificados, e estão descritos na literatura.

Tendo em conta que os doentes chegam por vezes tarde ao tratamento, é importante perceber quais são os factores que condicionam o atraso no tratamento, ou a chegada em tempo útil ao hospital, pelo que este trabalho tem como objectivo a identificação numa coorte, dos motivos pelos quais os doentes chegaram atrasados ao tratamento e os motivos que condicionaram a sua chegada mais precoce, após a instalação da sintomatologia de SCA.

A revisão da literatura científica permitiu contextualizar o problema, bem como sustentar as nossas hipóteses teóricas.

2.2 DEFINIÇÃO DA DOENÇA

Para o desenvolvimento deste trabalho torna-se essencial efectuar uma revisão da literatura, de forma a abranger temáticas relevantes para esta investigação, e também, para permitir a posterior interpretação e discussão dos resultados obtidos. Como tal, será feita uma breve exposição sobre alguns conceitos básicos tal como o conceito de miocárdio, uma sucinta explicação do procedimento de revascularização coronária, assim como a indicação das recomendações internacionais para os tempos de actuação terapêutica.

O Miocárdio é definido como a espessa camada média da parede cardíaca, composto de células musculares cardíacas, que são responsáveis pela capacidade contráctil do coração. Este músculo recebe todo o oxigénio necessário através das artérias coronárias[10].

A falta de oxigénio no músculo cardíaco, denominada por isquémia cardíaca, é secundária à perfusão inadequada do miocárdio, que gera desequilíbrio entre a oferta e a necessidade de oxigénio. A causa mais comum de isquémia miocárdica ou cardiopatia isquémica (CI) é a doença aterosclerótica obstrutiva das artérias coronárias[3, 11-12], que consiste numa acumulação de lípidos, hidratos de carbono complexos, sangue e seus produtos, tecido fibroso e depósitos de cálcio, na camada mais interna das artérias[4].

A redução da morbilidade e da mortalidade provocados pela CI, que é considerada responsável por mais mortes e incapacidade, acarretando maiores custos económicos do que qualquer outra patologia[3-4], é consequente de duas actuações, a prevenção por um lado e a optimização do tratamento das situações agudas por outro[13]. No entanto, ao longo da última década, apesar das medidas instituídas para prevenção e melhoria do tratamento, a prevalência e a incidência desta doença têm aumentado[13], sendo o SCA uma das mais sérias manifestações de doença cardíaca isquémica[14].

A denominação de SCA constitui uma terminologia recente que engloba muitos subtipos da cardiopatia isquémica aguda. Esta denominação engloba todo o espectro de manifestações patológicas associadas à CI e reflecte com maior precisão, a incerteza diagnostica que existe no momento da hospitalização, realçando o carácter urgente do problema e a sua localização e proporcionando um ponto de partida para uma série de decisões que determinam, rapidamente, o tratamento ideal e o diagnóstico definitivo[13]. Consoante as alterações que os doentes com SCA apresentavam no Electrocardiograma (ECG) e a presença de biomarcadores, estes são classificados em subgrupos, pelo que temos os doentes com Enfarte Agudo do Miocárdio com elevação de segmento ST no ECG (EAMEST), e os que possuem Enfarte Agudo do Miocárdio sem elevação de segmento ST no ECG (EAMSEST), e os que têm sintomas transitórios graves, cujo diagnóstico é a Angina Instável (AI)[3, 11-12].

A definição de EAM reflecte a morte celular das células miocárdicas causada por isquémia, este pode ser a primeira manifestação de doença coronária ou pode ocorrer, repetidamente, em pacientes com doença estabelecida.

O mecanismo fisiopatológico encontrado para o EAMEST afirma que este fenómeno ocorre quando o fluxo sanguíneo coronário diminui abruptamente depois de obstrução trombótica de uma artéria coronária previamente afectada pelo processo aterosclerótico[3, 12, 15].

O diagnóstico de SCA exige uma história clínica, um exame objectivo cuidadoso e um ECG de 12 derivações, em repouso. É útil proceder ao registo electrocardiográfico tanto durante a sintomatologia como após o seu desaparecimento. Com o desenvolvimento da tecnologia, existem hoje técnicas que associadas à história clínica e ao ECG permitem o diagnóstico de SCA, como é o caso dos biomarcadores específicos de necrose miocárdica, que correspondem a proteínas que são libertadas para o sangue a partir dos miócitos (células do músculo cardíaco) lesados e que são detectadas por testes sanguíneos que permitem, como o próprio nome indica, a identificação da necrose/lesão miocárdica[4].

Embora o tratamento básico dos pacientes com SCA seja clínico muitos indivíduos beneficiam dos procedimentos de revascularização coronária, que têm como fim restituir o fluxo sanguíneo às artérias coronárias, como é o caso da IPC. Outras técnicas de reperfusão podem ser utilizadas como é o caso da fibrinólise ou a cirurgia de *bypass* aorto-coronário[3]. No entanto as recentes *guidelines* recomendam a ICP primária como estratégia de reperfusão, quando realizada por operadores com experiência e no tempo recomendado[16].

A ICP consiste na introdução de cateteres na circulação arterial através da punção com uma agulha, os cateteres são avançados até ao coração sob orientação fluroscópica (injecção de contraste). O cateter é colocado no ostium da artéria coronária estenosada, permitindo a passagem do fio guia, dirigível e flexível, até a parte distal da artéria. Sobre este fio guia desloca-se o balão de angioplastia que ao ser insuflado irá aumentar o diâmetro da artéria estenosada (obstruída)[3].

A ICP pode então ser definida como primária quando a angioplastia é realizada sem fibrinólise prévia, que corresponde a um tratamento com fármacos e ICP com recurso quando é realizada após fibrinólise prévia. Nos casos agudos, é possível proceder a uma ICP da lesão implicada com uma taxa de sucesso superior a 95%[3]. Algumas das vantagens da ICP em comparação com a fibrinólise (terapêutica farmacológica do SCA) são: a redução da incidência de acidentes vasculares cerebrais hemorrágicos, menor incidência de reenfarte precoce, associada também ao facto de poder ser realizada quando existem contra-indicações para a terapêutica fibrinolítica.

Destacam-se como algumas vantagens da ICP em relação à revascularização miocárdica por bypass aorto-coronário, o facto de ser menos invasiva, condicionar uma hospitalização mais

curta, ter um custo inicial mais baixo, ser de fácil repetição, ser eficaz no alívio dos sintomas e o alívio da angina é alcançado na grande maioria dos casos[3].

Independentemente da técnica de reperfusão utilizada, o seu objectivo é minimizar o **tempo total de isquémia** (sofrimento do miocárdio), que para os pacientes com EAMEST é definido como o tempo desde o início da sintomatologia até ao início da terapia de reperfusão[7]. O EAMEST pode ser definido, no que diz respeito ao tempo, como "em desenvolvimento" quando o tempo desde o início dos sintomas até restabelecimento do fluxo é inferior a 6 horas, estudos demonstraram que a taxa de sobrevivência diminui drasticamente após as 6 horas, como se pode observar na Figura 1. Por esta razão a maior parte dos autores utilizam as 6 horas como referência para o melhor e o pior prognóstico[4, 17-18].

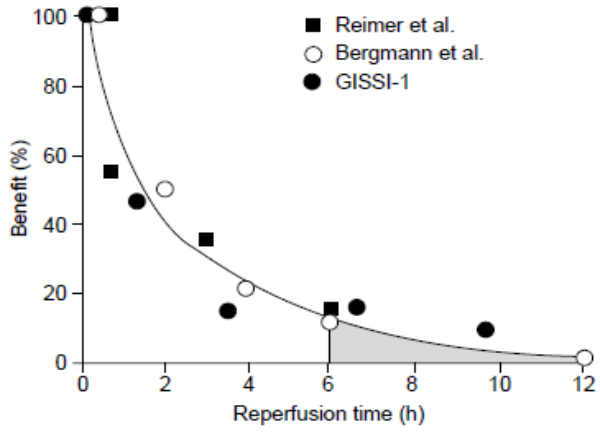

Figura 1: Benefício versus tempo de reperfusão a partir da combinação de vários estudos, Reimer et al, Bergmann et al e GISSI-1. A percentagem de benefício diz respeito a percentagem de miocárdio recuperado assim como a percentagem de redução da mortalidade[18].

Alguns estudos demonstram claramente haver uma melhoria dos resultados clínicos dos doentes que apresentam EAMEST com a ICP precoce[3, 13], uma vez que os benefícios que os doentes podem obter são "tempo-dependentes" a ICP primária deve ser realizada em carácter de emergência nas primeiras horas do enfarte[19-20].

Foi recentemente estabelecido que a reperfusão coronária alcançada por ICP de emergência pode reduzir a taxa de mortalidade hospitalar quando realizada o mais rapidamente possível[3]. Infelizmente, tem sido verificado que ao longo do tempo apenas uma pequena

percentagem dos pacientes com EAMEST realizam ICP primária dentro do tempo estipulado[21].

O facto de alguns estudos demonstrarem que as taxas de mortalidade aumentam com os aumentos no tempo de reperfusão, é uma das causas que levaram a uma investigação intensa de forma a tornar a ICP mais disponível no menor tempo possível[21-22].

2.3 IMPORTÂNCIA DO ESTUDO

De todos os pacientes com doença coronária que morrem dentro de 28 dias após o início dos sintomas, cerca de dois terços morrem antes de dar entrada no hospital. Este facto destaca a necessidade de investigação de forma a identificar os factores associados com o aumento do tempo total de isquémia, a necessidade de reconhecimento precoce dos sinais de alerta de um SCA, mas também a necessidade de prevenção e educação da população geral[2].

De acordo com o registo nacional de síndromes coronários agudos da Sociedade Portuguesa de Cardiologia, foram registados 5384 novos casos no ano de 2003 e 3834 no ano de 2004. Relativamente à evolução da mortalidade intra-hospitalar registaram-se 6.3% de mortes no EAMEST, 3.3% no EAMSEST e 0.4% na AI[6].

Estudos demonstram que o SCA comporta elevadas taxas de mortalidade nas primeiras horas de evolução, podendo até atingir taxas de mortalidade de 50% na fase pré-hospitalar para o EAMEST[23].

Assim, uma vez que o SCA é a principal causa de mortalidade e morbilidade nos países desenvolvidos, torna-se pertinente seguir uma linha de investigação que permita verificar se o tempo entre o início dos sintomas e o restabelecimento do fluxo, isto é o tempo total de isquémia, por angiografia coronária se encontra dentro do recomendado pelas *guidelines* ou se, pelo contrário, este é excessivamente demorado, ultrapassando o recomendado nas *guidelines*, o que pode ser um factor contributivo para o elevado número de mortes e incapacidades[3-4].

Na maioria dos estudos foi demonstrado que um aumento do tempo total de isquémia estava associado a um pior prognóstico[24], avaliado através do tamanho da zona de enfarte, assim como da taxa de mortalidade. Numa análise recente foi comprovado que por cada 30 minutos

de atraso no tempo de reperfusão estava associado a um aumento do risco relativo de morte de 7.5%, por um ano[25].

Capítulo 3

Modelo Linear Generalizado:

Modelo de Regressão Logística

3.1 INTRODUÇÃO

Em investigação clínica é frequente a variável resposta ser binária resultante por exemplo, da presença ou ausência de determinado sintoma, como é o caso dos dados analisados neste trabalho. O modelo de regressão logística é a metodologia estatística adequada para a análise deste tipo de dados. Este modelo insere-se nos chamados modelos lineares generalizados (MLGs) que foram apresentados pela primeira vez num artigo de Nelder e Wedderburn em 1972 e que são uma extensão dos modelos lineares.

Tendo em atenção o que se acaba de expôr vai-se por começar por apresentar os MLGs de uma forma geral passando-se de seguida para o caso particular do modelo de regressão logística.

3.2 MODELO LINEAR GENERALIZADO

A origem dos MLGs reside nos avanços do conhecimento estatístico assim como no rápido desenvolvimento computacional, estes modelos correspondem a uma síntese dos modelos lineares clássicos (MLs) e de outros modelos, tendo sido unificada, quer do ponto de vista teórico quer do ponto de vista conceptual, a teoria da modelação estatística até então desenvolvida. A ideia principal é abrir um leque de opções para a distribuição da variável resposta, possibilitando que a mesma pertença à família exponencial de distribuições, bem como dar maior flexibilidade à relação entre o valor médio da variável resposta e o preditor linear. A ligação entre o valor médio e o preditor linear pode assumir qualquer forma monótona não-linear, não sendo necessariamente a identidade[26-27].

Os MLGs abrangem uma grande classe de modelos estatísticos, todos com o objectivo de relacionar a variável resposta com a combinação linear de variáveis explicativas[1]. Estes modelos permitem, para além de modelos de regressão para variáveis resposta contínuas, modelos de regressão para taxas e proporções, para dados binários, para dados ordinais, para variáveis multinomiais e contagens, entre outros.

Uma abordagem feita através dos MLG oferece várias vantagens, nomeadamente:

(1) fornece uma estrutura teórica geral para a maioria dos modelos estatísticos usados na prática;
(2) simplifica a implementação destes diferentes modelos nos diferentes *softwares* estatísticos, uma vez que, essencialmente, o mesmo algoritmo pode ser utilizado para a estimação, inferência e avaliação da adequação do modelo para todos os MLGs.

Esta generalização é obtida estendendo as hipóteses subjacentes ao MLs em duas direcções:

1. Variáveis resposta com outras distribuições que não a distribuição normal.
2. Relação entre a resposta e as variáveis explicativas estabelecida por outra função de ligação que não a linear, dependendo do tipo de resposta que está a ser analisada.

Os MLGs são portanto uma extensão dos modelos lineares que englobam os modelos com variável resposta de distribuição não normal[27]. Esta extensão só foi alcançada após o reconhecimento de que muitas das propriedades da distribuição normal eram também partilhadas pelas distribuições de família exponencial, nas quais se encontram incluídas distribuições como a Bernoulli, a binomial, a Poisson, a exponencial, a gama, a binomial negativa, a multinomial[26].

[1] Ao longo deste trabalho usar-se-á, indiferentemente, variável explicativa ou covariável.

3.2.1 A FAMÍLIA EXPONENCIAL

Designe-se por $\mathbf{Y} = (Y_1,...,Y_n)^T$ o vector aleatório constituído por *n* variáveis aleatórias (v.a´s) independentes. Diz-se que cada component Y_i tem uma distribuição pertencente à família exponencial e escreve-se $Y_i \sim FE\left(b(\theta_i), \frac{\phi}{\omega_i}\right)$ se a função densidade probabilidade (f.d.p.) ou a função massa probabilidade (f.m.p.) assumir a forma:

$$f\left(y_i \mid \theta_i, \omega_i, \phi\right) = \exp\left\{\frac{\omega_i}{\phi}\left[y_i\theta_i - b\left(\theta_i\right)\right] + c\left(y_i, \phi\right)\right\} \tag{3.1}$$

onde θ e ϕ são parâmetros escalares, $b(.)$ e $c(.)$ são funções reais conhecidas e ω_i é uma constante conhecida que varia de observação para observação e à qual se dá o nome de peso. Na definição θ é a **forma canónica do parâmetro de localização** e ao parâmetro ϕ, em geral conhecido, dá-se o nome de **parâmetro de dispersão** ou de escala, sendo constante ao longo das observações. Assume-se ainda que a função $b(.)$ é diferenciável e que o suporte da distribuição não depende de parâmetros desconhecidos.

Para qualquer escolha do parâmetro de dispersão, ϕ, temos uma família exponencial no entanto, se θ e ϕ variarem simultaneamente, pode não se ter uma família exponencial. Quando o parâmetro ϕ é desconhecido a distribuição pode ou não fazer parte da família exponencial (*vide in* página 5, *Modelos Lineares Generalizados-da teoria à prática*[28]).
Nas circunstâncias acima apresentadas, a família exponencial obedece às condições habituais de regularidade[29-31].

Prova-se que se Y_i tem uma distribuição pertencente à família exponencial então[28]:

$$E(Y_i) = b'(\theta_i) \tag{3.2}$$

e

$$Var(Y_i) = \frac{\phi}{\omega_i} b''(\theta_i) \tag{3.3}$$

Tem-se assim que a variância de Y_i depende da função $b''(\theta_i)$ que depende do parâmetro canónico (ou seja depende do valor médio). A esta função dá-se o nome de *função de variação* e será designada por $V(\mu_i)$, donde $V(\mu_i) = b''(\theta_i)$.

Exemplo – Seja $Y \sim N(\mu, \sigma^2)$**.** A f.d.p é dada por $f(y \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$ que pode assumir a forma $exp\left\{\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right) - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right)\right\}$ para $y \in \Re$ ou seja, $Y \sim FE\left(b(\theta), \frac{\phi}{\omega}\right)$ com $\theta = \mu$, $b(\theta) = \frac{\mu^2}{2}$, $c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right), \phi = \sigma^2$ e $\omega = 1$.

Tem-se ainda $E(Y) = \mu = b'(\theta)$ e $Var(Y) = \frac{\phi}{\omega} b''(\theta) = \sigma^2$

3.2.2 EXTENSÃO DO ML AO MLG

Considerem-se *n* observações independentes realização da v.a Y a que se dá o nome de variável resposta e seja Y_i, $i = 1,...,n$, a variável resposta para o *i*-ésimo indivíduo e $\mathbf{y} = (y_1,...,y_n)^T$ o vector de observações, em que y_i é a observação para o *i*-ésimo indivíduo. Associado a cada variável resposta, Y_i, tem-se o vector $p \times 1$ de covariáveis, $\mathbf{x}_i^T = (x_{i1},...,x_{ip})$, $i = 1,...,n$, onde x_{ik} representa a *k*-ésima covariável para o *i*-ésimo indivíduo, e $\boldsymbol{\beta} = (\beta_1,..,\beta_p)^T$ um vector $p \times 1$ de parâmetros desconhecidos sendo, na maioria dos casos, $x_{i1} = 1$ para qualquer *i*.

A parte sistemática (ou determinística) do ML escrever-se-á:

$$\mu_i = \sum_{j=1}^{p} x_{ij}\beta_j = \mathbf{x}_i^T \boldsymbol{\beta} \qquad (3.4)$$

$i = 1,...,n$

onde x_{ij} é o valor da *j*-ésima covariável para a observação *i* e β_j o *j*-ésimo parâmetro desconhecido.

Ao definir-se o **preditor linear** por $\eta_i = \mathbf{x_i^T \beta}$ torna-se pois evidente que, a relação entre o valor médio, μ_i, e o preditor linear, η_i, é a identidade.

O ML pode ser especificado em três partes, a saber:

1. a **componente aleatória**: as variáveis aleatórias Y_i são independentes e têm distribuição normal com $E(Y_i) = \mu_i$ e variância constante, σ^2, ou seja,

$$Y_i \sim N(\mu_i, \sigma^2) \qquad i = 1,\ldots,n;$$

2. a **componente sistemática**: um preditor linear dado por:

$$\eta_i = \mathbf{x_i^T \beta};$$

3. a **ligação** entre a componente aleatória e sistemática:

$$\mu_i = \eta_i .$$

Os MLG são obtidos estendendo as hipóteses subjacentes ao ML em duas direcções:

1. a distribuição de Y_i pode ser uma qualquer da família exponencial, tal que $b'(\theta_i) = \mu_i$ é permitido;

2. outras formas de ligação, além da identidade, entre o preditor linear, η_i, e o valor médio, μ_i, são possíveis, isto é,

$$g(\mu_i) = \eta_i$$

onde $g(.)$ é uma função monótona diferenciável à qual se dá o nome de **função de ligação.**

Tem-se assim que os MLGs são assim caracterizados pela seguinte estrutura:

1. **componente aleatória**: dado o vector de covariáveis $\mathbf{x}_i$ as variáveis aleatórias Y_i são condicionalmente independentes com distribuição pertencente à família exponencial;

2. **componente sistemática**: define-se um preditor linear η_i dado por variáveis explicativas

$$\eta_i = \mathbf{x_i^T \beta};$$

3. **função de ligação** entre a componente aleatória e sistemática dada por:

$$g(\mu_i) = \eta_i .$$

3.2.3 MÉTODOS DE ESTIMAÇÃO

Tal como para qualquer outro modelo linear, para alcançar o objectivo de descrever a relação entre a variável resposta e as variáveis explicativas, é essencial a estimação dos parâmetros desconhecidos do modelo. No MLG o parâmetro de interesse é $\boldsymbol{\beta}$ sendo a sua estimação baseada no método da máxima verosimilhança. O parâmetro de dispersão ϕ, quando existe, é considerado parâmetro perturbador e é estimado pelo método dos momentos[28].

A função de verossimilhança de um MLG com respostas independentes e supondo que ϕ conhecido, é dada por [28]:

$$L = \prod_{i=1}^{n} L_i(\theta_i, \omega_i, \phi) = \prod_{i=1}^{n} f(y_i \mid \theta_i \omega_i, \phi) = \exp\left[\sum_{i=1}^{n} \frac{\omega_i}{\phi} \left(y_i \theta_i - b(\theta_i) \right) + \sum_{i=1}^{n} c(y_i, \phi) \right] \quad (3.5)$$

Sendo o logaritmo da função verosimilhança $l = \log L$

$$l = \sum_{i=1}^{n} l_i(\theta_i, \omega_i, \phi) = \sum_{i=1}^{n} \frac{\omega_i}{\phi} \left(y_i \theta_i - b(\theta_i) \right) + \sum_{i=1}^{n} c(y_i, \phi) \quad (3.6)$$

As equações de máxima verosimilhança para $\boldsymbol{\beta}$, com base numa amostra aleatória de dimensão *n* e nas condições anteriormente mencionadas para o modelo, são dadas por:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = 0, \quad j = 1, \ldots, p \quad (3.7)$$

Como l é função composta dos β_j, $j = 1, \ldots, p$, vem

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \quad (3.8)$$

Tendo em atenção que $b'(\theta_i) = \mu_i$ e $b''(\theta_i) = \frac{\omega_i Var(Y_i)}{\phi}$, então:

$$\begin{cases} \dfrac{\partial l_i}{\partial \theta_i} = \dfrac{\omega_i (y_i - \mu_i)}{\phi} \\ \dfrac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \dfrac{\omega_i Var(Y_i)}{\phi} \\ \dfrac{\partial \eta_i}{\partial \beta_j} = x_{ij} \end{cases} \tag{3.9}$$

pelo que a equação dada em (3.8) pode ser escrita na seguinte forma:

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{(y_i - \mu_i) x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \tag{3.10}$$

Finalmente as equações de máxima verosimilhança para $\boldsymbol{\beta}$ assumem a forma

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \qquad j = 1, \ldots, p. \tag{3.11}$$

a U_j dá-se o nome de score e a $\mathbf{U} = (U_1, \ldots, U_p)^T$ dá-se o nome de função score.

Como antes referido, o conhecimento do parâmetro de dispersão, para a estimação de $\boldsymbol{\beta}$, é irrelevante.

A matriz de variância-covariância da função *score* é designada por matriz de informação de Fisher e é dada por:

$$I(\boldsymbol{\beta}) = E\left[-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right] \tag{3.12}$$

é obtida considerando as segundas derivadas de l_i e os seus valores esperados. Para famílias regulares tem-se que:

$$-E\left[\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} \right] = E\left[\frac{\partial l_i}{\partial \beta_j} \frac{\partial l_i}{\partial \beta_k} \right] = \frac{x_{ij} x_{ik}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \tag{3.13}$$

O (j,k) - ésimo elemento da matriz $I(\boldsymbol{\beta})$ é dado por:

$$-\sum_{i=1}^{n} E\left[\frac{\partial^2 l_i}{\partial\beta_j\partial\beta_k}\right] = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{Var\left(Y_i\right)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \tag{3.14}$$

Em notação matricial tem-se:

$$I\left(\boldsymbol{\beta}\right) = \mathbf{X}^{\mathbf{T}}\tilde{\mathbf{W}}\mathbf{X}$$

onde $\tilde{\mathbf{W}}$

$$\tilde{\mathbf{W}} = \begin{bmatrix} \tilde{\omega}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\omega}_2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \\ 0 & 0 & \cdots & \tilde{\omega}_n \end{bmatrix} \tag{3.15}$$

e

$$\tilde{\omega}_i = \frac{1}{Var\left(Y_i\right)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \tag{3.16}$$

Todos os cálculos algébricos estão apresentados com mais detalhe em McCullagh & Nelder (1989)[32], Azzalini (1996)[30] ou em Dobson (2002)[26].

As equações de máxima verosimilhança referidas em (3.11), não têm solução analítica, pelo que é necessário recorrer a métodos iterativos. A única excepção é o caso em que a distribuição da variável resposta é normal. Nesta situação o método da máxima verosimilhança é precisamente o método dos mínimos quadrados ponderados.

O termo "mínimos quadrados ponderados" é usado no sentido em que os cálculos computacionais envolvem funções não lineares da resposta e o vector de pesos é recalculado em cada iteração [33]. Com base no **método iterativo dos mínimos quadrados ponderados** obtêm-se o estimador de máxima verosimilhança de $\boldsymbol{\beta}$ no preditor linear, η [30, 32](McCullagh & Nelder, 1989; e Azzalini,1996). Para descrição deste algoritmo ver Gonçalves (2002).

Apesar do **parâmetro de dispersão** ou escala, ϕ, poder ser estimado através do método de máxima verosimilhança, existe um método mais simples que dá geralmente bons resultados.

Este método é baseado na distribuição de amostragem da estatística de Pearson generalizada, para valores de *n* suficientemente grande, sendo o estimador de ϕ dado por[28]:

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{\omega_i\left(Y_i-\hat{\mu}_i\right)^2}{V\left(\hat{\mu}_i\right)} \tag{3.17}$$

3.2.4 INFERÊNCIA NO MODELO LINEAR GENERALIZADO

Depois de obter as estimativas para os coeficientes interessa avaliar a significância desses coeficientes. Este passo envolve geralmente a formulação de testes de hipóteses e construção de intervalos de confiança de modo a determinar se as variáveis introduzidas no modelo estão significativamente associadas à variável resposta[34]. A inferência, baseada quer nos testes quer nos intervalos de confiança, só é possível sabendo a distribuição das respectivas estatísticas o que requer o conhecimento da distribuição dos estimadores de máxima verosimilhança de $\boldsymbol{\beta}$ e das suas propriedades.

Distribuições Assimptóticas

Começar-se-á por enunciar a distribuição assimptótica do estimador de máxima verosimilhança de $\boldsymbol{\beta}$ e as suas propriedades. A justificação pode ser encontrada na referência[28].

1. O estimador de máxima verosimilhança de $\boldsymbol{\beta}$ é assimptoticamente centrado sendo a matriz de variância-covariância aproximadamente igual a $I^{-1}(\boldsymbol{\beta})$.
2. A distribuição assimptótica de $\hat{\boldsymbol{\beta}}$ é normal p-variada com valor médio $\boldsymbol{\beta}$ e matriz de variância-covariância $I^{-1}(\boldsymbol{\beta})$ e escreve-se $\hat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, I^{-1}(\boldsymbol{\beta})\right)$.
3. A $\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T I(\boldsymbol{\beta})\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$ dá-se o nome de estatística de Wald e a distribuição assimptótica é qui-quadrado com p graus de liberdade $\left(\chi_p^2\right)$.
4. A distribuição assimptótica de $\hat{\beta}_j,\ j=1,\ldots,p$ é normal com valor médio β_j e variância $I_{jj}^{-1}(\boldsymbol{\beta})$ e escreve-se $\hat{\beta}_j \sim N\left(\beta_j, I_{jj}^{-1}(\boldsymbol{\beta})\right)$ onde $I_{jj}^{-1}(\boldsymbol{\beta})$ é o elemento (j,j) de $I^{-1}(\boldsymbol{\beta})$.

Como $\boldsymbol{\beta}$ é desconhecido e a matriz de informação de Fisher depende de $\boldsymbol{\beta}$ esta é desconhecida pelo que se substitui $I^{-1}(\boldsymbol{\beta})$ por $I^{-1}(\hat{\boldsymbol{\beta}})$.

Testes de Hipóteses

A maior parte dos testes de hipótese sobre o vector $\boldsymbol{\beta}$, podem ser formulados em termos de hipóteses lineares da forma:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi} \quad vs \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\xi}$$

onde $\mathbf{C}$ é uma matriz $q \times p$, com $q \leq p$ de característica completa q, e $\boldsymbol{\xi}$ é um vector de dimensão q previamente especificado [28].

Casos especiais da hipótese anterior são:

- Hipótese da nulidade de uma componente do vector parâmetro, nomeadamente:

  $$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0,$$

  para algum j, sendo neste caso q=1, $\mathbf{C}$ =(0,….0,1,0…0) e ocupando o 1 a j-ésima posição e $\boldsymbol{\xi}$ =0.

- Hipótese da nulidade de r componentes de $\boldsymbol{\beta}$. Se tivermos por exemplo: $H_1 : (\beta_1 \ldots \beta_r)^T = (0, \ldots, 0)^T$, então $q = r$ e $\mathbf{C} = (I_r O_{r \times (p-r)})$ $\boldsymbol{\xi} = \mathbf{0}_r$

  onde I_r é a matriz identidade de dimensão r, $O_{r \times (p-r)}$ é uma matriz de zeros de dimensão $r \times (p-r)$ e $\mathbf{0}_r$ é o vector nulo de dimensão r.

Qualquer uma das hipóteses consideradas corresponde a testar submodelos do modelo original considerado. Mais especificamente, a primeira hipótese consiste em testar um submodelo com todas as covariáveis do modelo original à excepção da covariável x_j relativa ao parâmetro de regressão β_j e a segunda consiste em testar um modelo sem as r covariáveis relativas aos parâmetros supostos nulos na hipótese H_0.

Existem essencialmente três estatísticas para testar as hipóteses referidas sendo aqui abordadas as duas mais usadas: a **Estatística de Wald** e a **Estatística de Razão de Verosimilhanças ou Estatística de Wilks** dando origem aos testes com o mesmo nome.

Teste de Wald

A estatística de Wald é, em geral, utlizada para testar hipóteses nulas sobre as componentes individuais. Neste caso a estatística é dada por:

$$W = \left(\hat{\beta}_j - \beta_j\right)^T I_{jj}\left(\hat{\boldsymbol{\beta}}\right)\left(\hat{\beta}_j - \beta_j\right) \tag{3.18}$$

que sob a hipótese H_0 assume a forma $W = \dfrac{\hat{\beta}_j^2}{I_{jj}^{-1}\left(\hat{\boldsymbol{\beta}}\right)}$ e segue uma distribuição assimptótica de um χ_1^2. A hipótese nula é rejeitada ao nível de significância α se o valor observado da estatística de teste for superior ao quantil de probabilidade $1-\alpha$ do χ_1^2.

Em muitos programas estatísticos a distribuição indicada neste caso é a normal e a estatística de Wald é dada por:

$$\sqrt{W} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{I_{jj}^{-1}\left(\hat{\boldsymbol{\beta}}\right)}} \tag{3.19}$$

A estatística de Wald é igualmente utilizada para testar a nulidade de r componentes de $\boldsymbol{\beta}$, assumindo a forma:

$$W = \left(\mathbf{C}\hat{\boldsymbol{\beta}}\right)^T \left[\mathbf{C}I^{-1}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{C}^T\right]\left(\mathbf{C}\hat{\boldsymbol{\beta}}\right)$$

e, sob a hipótese H_0, segue uma distribuição assimptótica de um qui-quadrado com *r* graus de liberdade, χ_r^2. A hipótese nula é rejeitada ao nível de significância α se o valor observado da estatística de teste for superior ao quantil de probabilidade $1-\alpha$ do χ_r^2.

Teste de Razão de Verosimilhanças

A estatística de razão de verosimilhanças ou de Wilks é definida por:

$$\Lambda = 2\left\{\ell\left(\tilde{\boldsymbol{\beta}}\right) - \ell\left(\hat{\boldsymbol{\beta}}\right)\right\} = -2\left\{\ell\left(\hat{\boldsymbol{\beta}}\right) - \ell\left(\tilde{\boldsymbol{\beta}}\right)\right\}$$

onde $\tilde{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\beta}}$ são os estimadores de máxima verosimilhança de $\boldsymbol{\beta}$ sob $H_0 \cup H_1$ e H_0, respectivamente.

O teorema de Wilks (e.g., Cox e Hinkley, 1974) estabelece que sob certas condições de regularidade e sob H_0, Λ segue assimptoticamente uma distribuição qui-quadrado com graus de liberdade iguais à diferença entre o número de parâmetros a estimar sob $H_0 \cup H_1$ e o número de parâmetros a estimar sob H_0, ou seja *r*.

De acordo com este teste a hipótese nula é rejeitada a um nível de significância α, se o valor observado da estatística Λ for superior ao quantil de probabilidade $1-\alpha$ do χ^2_r.

Intervalos de confiança

Os intervalos de confiança (IC) para os parâmetros β_j, $j = 1,\ldots,p$ ao nível de confiança $(1-\alpha)$ podem ser obtidos com base na estatística de Wald, através seguinte expressão:

$$\beta_j \pm z_{1-\alpha/2}\, SE(\hat{\beta}_j)$$

onde $z_{1-\alpha/2}$ é o quantil de $(1-\alpha/2)$ para a distribuição normal padrão e $SE\left(\hat{\beta}_j\right) = \sqrt{I^{-1}_{jj}\left(\hat{\boldsymbol{\beta}}\right)}$

Para o vector $\boldsymbol{\beta}$ de dimensão *p*:

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T I\left(\hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \leq \chi^2_{(1-\alpha,p)}$$

onde $\chi_{(1-\alpha,p)}$ é o quantil de $(1-\alpha)$ de um χ^2 com *p* graus de liberdade dá-nos o elipsoide de confiança para $\boldsymbol{\beta}$.

3.2.5 SELECÇÃO DO MODELO

O problema da selecção do modelo corresponde à procura do melhor modelo, isto é, saber qual é o modelo mais parcimonioso. Dito de outro modo identificar o modelo que, com o

menor número de covariáveis, consegue ajustar-se bem aos dados e ainda oferecer uma boa interpretação do problema em estudo.

Durante o processo de selecção existe uma série de modelos em consideração dos quais os habitualmente utilizados são:

- Modelo Saturado para um MLG com n observações é o modelo com o número máximo de parâmetros, isto é, com n parâmetros (um para cada observação) e como tal o modelo ajusta-se perfeitamente aos dados. Neste modelo toda a variação é atribuída à componente sistemática do modelo.

- Modelo Nulo é o modelo que contém um único parâmetro, assume que todas as variáveis Y_i têm o mesmo valor médio μ. Toda a variação do modelo é atribuída à componente aleatória.

- Modelo maximal é o modelo que contém o maior número de parâmetros sendo portanto o modelo mais complexo que se irá considerar.

- Modelo minimal é o modelo mais simples, ou seja, com o menor número de parâmetros que ainda se ajusta adequadamente aos dados. Este modelo pode no entanto esconder características presentes nos dados.

- Modelo corrente é qualquer modelo com q parâmetros linearmente independentes, situado entre o modelo maximal e o modelo minimal e que está a ser analisado.

Estatísticas para a selecção do modelo

Quando se pretende decidir entre dois modelos qual o que deve ser rejeitado ou não duas situações podem ocorrer:

(i) Os modelos estão encaixados;

(ii) Os modelos não estão encaixados.

No primeiro caso (i) quer o teste de razão de verosimilhanças quer o teste de Wald, descritos na secção anterior, podem ser utilizados. No segundo caso (ii) utilizam-se os chamados:

Critérios de informação

Quando os modelos não estão encaixados um critério de selecção que pode ser aplicado é o Critério de Informação de Akaike (AIC) que, para um modelo com *p* parâmetros, é dado por:

$$AIC = Desvio\ (mod\,elo) + 2p$$

Um outro critério igualmente utilizado é o Critério de Informação Bayesiano (BIC) que, para um modelo com p parâmetros é dado por:

$$BIC = Desvio(modelo) + 2p\log(n)$$

Qualquer um destes critérios é baseado na função log-verosimilhança com um factor de penalização para o número de parâmetros.

Quanto menor o valor obtido para o AIC ou para o BIC melhor será o modelo em investigação.

Função desvio e comparação de modelos encaixados

Sejam $\tilde{\boldsymbol{\beta}}_S$ e $\hat{\boldsymbol{\beta}}_M$ os estimadores de máxima verosimilhança de $\boldsymbol{\beta}$ para o modelo saturado, S, e para o modelo corrente M. A estatística de teste de razão de verosimilhanças tal como foi definida na secção anterior é dada por:

$$\Lambda = 2\left\{\ell\left(\tilde{\boldsymbol{\beta}}_S\right) - \ell\left(\hat{\boldsymbol{\beta}}_M\right)\right\} = -2\left\{\ell\left(\hat{\boldsymbol{\beta}}_M\right) - \ell\left(\tilde{\boldsymbol{\beta}}_S\right)\right\} \qquad (3.20)$$

e pode ser escrita na forma:

$$-2\left\{\ell\left(\hat{\boldsymbol{\beta}}_M\right) - \ell\left(\tilde{\boldsymbol{\beta}}_S\right)\right\} = -2\sum_i \frac{\omega_i}{\phi}\left\{\left[y_i\hat{\theta}_i - b\left(\hat{\theta}_i\right)\right] - \left[y_i\tilde{\theta}_i - b\left(\tilde{\theta}_i\right)\right]\right\} = \frac{D(\mathbf{y};\hat{\boldsymbol{\mu}})}{\phi} \qquad (3.21)$$

onde $\hat{\theta}_i$ e $\tilde{\theta}_i$ são os estimadores de máxima verosimilhança de θ_i para os modelos M e S, respectivamente.

A $\frac{D(\mathbf{y};\hat{\boldsymbol{\mu}})}{\phi}$ dá-se o nome de **desvio reduzido** e ao seu numerador $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ **desvio** para o modelo corrente.

O desvio $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ pode ainda ser escrito na forma:

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = \sum_i 2\omega_i \left\{ \left[y_i \left(\tilde{\theta}_i - \hat{\theta}_i \right) - b\left(\tilde{\theta}_i \right) + b\left(\hat{\theta}_i \right) \right] \right\} = \sum_i d_i \qquad (3.22)$$

onde d_i mede a diferença dos logaritmos das verosimilhanças observada e ajustada para a i-ésima observação, sendo só função dos dados.

Uma propriedade importante do desvio é a aditividade para modelos encaixados.

Para dois modelos intermédios M_m e Mq com p_m e p_q parâmetros, respectivamente, e tais que $M_m \subset Mq$ a estatística da razão de verosimilhanças para comparar estes dois modelos é dada por:

$$-2\left\{ \ell_{Mm}\left(\hat{\boldsymbol{\beta}}\right) - \ell_{Mq}\left(\hat{\boldsymbol{\beta}}\right) \right\} = -2\left[\ell_{Mm}\left(\hat{\boldsymbol{\beta}}\right) - \ell_S\left(\hat{\boldsymbol{\beta}}\right) \right] + 2\left[\ell_{Mq}\left(\hat{\boldsymbol{\beta}}\right) - \ell_S\left(\hat{\boldsymbol{\beta}}\right) \right] = \frac{Dm - Dq}{\phi} \quad (3.23)$$

onde D_m e D_q designam os desvios dos modelos. Com base no que foi dito na secção anterior, sob a hipótese do modelo Mm ser o correcto segue uma distribuição assimptótica de um $\chi^2_{p_q - p_m}$. Tem-se assim que a comparação de modelos encaixados pode ser feita através da diferença dos desvios de cada modelo.

3.3 MODELO DE REGRESSÃO LOGÍSTICA

Considere-se agora que a variável resposta Y é binária, isto é, assume o valor 1 ou 0 consoante se observa a presença $(Y=1)$ ou ausência $(Y=0)$ de determinado sintoma. Assumindo que $P(Y=1)=\pi$ é a probabilidade de sucesso, a variável resposta Y segue então uma distribuição de Bernoulli de parâmetro π.

Em muitas situações experimentais as respostas aparecem agrupadas sobre a forma de proporções resultantes do facto de mais de um indivíduo partilhar a mesma combinação de condições experimentais. Assim, se m_i for o número de réplicas (indivíduos) para cada combinação $\mathbf{x}_i^T = (x_{i1},...,x_{ip})$ das covariáveis então a v.a. $\tilde{Y}_i$ correspondente ao número de sucessos é binomial de parâmetros m_i e π_i.

Tendo em atenção que o objectivo é o de investigar a relação entre a probabilidade da resposta, π_i , e as variáveis explicativas $\mathbf{x}_i^T = (x_{i1},...,x_{ip})$ não se irá usar $\tilde{Y}_i$ como a variável resposta mas sim a proporção de sucessos $Y_i = \frac{\tilde{Y}_i}{m_i}$. Deste modo $E(Y_i) = \pi_i$ é, de facto, a probabilidade pretendida (*vide in* página 51 Collett)[35].

A f.m.p da v.a. Y é dada por (omitiu-se o índice i):

$$
\begin{aligned}
f(y) &= P(Y=y) = \binom{m}{my}\pi^{my}(1-\pi)^{m-my} = \binom{m}{my}\left(\frac{\pi}{1-\pi}\right)^{my}(1-\pi)^{m} \\
&= \binom{m}{my}\left(\frac{\pi}{1-\pi}\right)^{my}\left(1+\frac{\pi}{1-\pi}\right)^{-m} = \exp\left\{my\log\left(\frac{\pi}{1-\pi}\right) - m\log\left(1+\frac{\pi}{1-\pi}\right) + \log\binom{m}{my}\right\} \qquad (3.24) \\
&= \exp\left\{my\log\left(\frac{\pi}{1-\pi}\right) - m\log\left[1+\exp\left(\log\left(\frac{\pi}{1-\pi}\right)\right)\right] + \log\binom{m}{my}\right\} \\
&= \exp\left\{m\left[y\theta - b(\theta)\right] + c(y,\phi)\right\},
\end{aligned}
$$

com $\theta = \log\left(\frac{\pi}{1-\pi}\right), b(\theta) = log[1+exp(\theta)],\ \phi = 1,\ \omega = m$ e $c(y,\phi) = \log\binom{m}{my}$

para $y = 0, \frac{1}{m}, \frac{2}{m}, .., 1,$ pertencendo assim à família exponencial e tem-se:

$$
E(Y) = b'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \frac{\frac{\pi}{1-\pi}}{\frac{1}{1-\pi}} = \pi \text{ e } \operatorname{Var}(Y) = \frac{\phi}{\omega}b''(\theta) = \frac{1}{m}\frac{e^{\theta}}{(1+e^{\theta})^2} = \frac{1}{m}\frac{\left(\frac{\pi}{1-\pi}\right)}{\frac{1}{(1-\pi)^2}} = \frac{\pi(1-\pi)}{m}.
$$

O parâmetro canónico é a função *logit,* $\log\left(\frac{\pi}{1-\pi}\right)$.

Quando $m_1 = ... = m_n = 1$ diz-se que se tem **dados binários não agrupados**.

Se admitirmos que a relação entre π_i e as covariáveis $\mathbf{x}_i^T = (x_{i1},...x_{ip})$ é linear está-se perante um modelo linear generalizado cuja parte determinística é dada por:

$$
g(\pi_i) = \eta_i = \mathbf{x_i^T\beta}
$$

Para se saber qual a função de ligação a considerar basta ter em atenção que $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$

pelo que se conclui que a função de ligação canónica é o *logi*[2]*t* o que leva ao **modelo de regressão logística** quer tenhamos dados binários agrupados ou não.

A parte determinística do modelo de regressão logística é dado por:

$$logit\left(\pi_i\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta} \tag{3.25}$$

onde π_i é a probabilidade de sucesso, $\pi_i = P\left(Y=1 \mid \mathbf{x}_i^T\right)$. A transformação alcançada pela função de ligação *logit* permite obter as propriedades desejáveis do modelo linear de regressão[36].

Possui ainda vantagens como a de ser particularmente apropriada para dados provenientes de estudos retrospectivos e permitir estimar diferenças na escala *logit* quer os dados sejam provenientes de estudos retrospectivos ou de estudos prospectivos[34].

A probabilidade de sucesso, π_i, é obtida através de:

$$\pi_i = \frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)} \tag{3.26}$$

3.3.1 INTERPRETAÇÃO DOS PARÂMETROS DO MODELO

Duas noções que assumem particular importância na interpretação dos modelos logísticos são a de *odds* e *odds ratio* (OR).

[2] Existem outras funções de ligação[36. Felgueiras, M.M., *Análise de dados binários*, in *Faculdade de Ciências*. 2003, Universidade de Lisboa: Lisboa.] mas não serão aqui abordadas.

Definição 3.1 - Chama-se *odds* de um acontecimento, ao quociente entre a probabilidade de sucesso desse acontecimento, definido como π_i, e a probabilidade de insucesso, $1-\pi_i$, isto é:

$$Odds = \frac{\pi_i}{1-\pi_i}$$

O *odds* ao contrário da probabilidade pode assumir qualquer valor positivo.

Definição 3.2 - Quando dois conjuntos $(i=1 \text{ e } i=2)$ de dados binários são comparados dá-se o nome de OR ao quociente:

$$OR = \frac{\pi_1 / (1-\pi_1)}{\pi_2 / (1-\pi_2)}$$

1. Quando os *odds* nos dois conjuntos de dados binários são iguais o *OR* é igual a 1. Isto acontece quando as probabilidades de sucesso são iguais.

2. Valores de *OR* menores do que 1 sugerem que o *odds* no primeiro conjunto é menor do que no segundo. Valores de *OR* maiores do que 1 sugerem a situação contrária.
3. O *odds ratio* é uma medida da diferença entre duas probabilidades de sucesso que pode tomar qualquer valor positivo, ao contrário de $\pi_1 - \pi_2$ que varia no intervalo $(-1,1)$.

4. Se considerarmos o logaritmo tem-se:

$$\log OR = \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_2}{1-\pi_2}\right)$$

que não mais é do que a diferença dos *logit* nos dois conjuntos.

5. Os *odds ratio* podem descrever o efeito do tratamento independentemente das covariáveis.

Seja:

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

onde o índice *i* foi omitido e $\pi = \pi(\mathbf{x}) = P(Y = 1 \mid \mathbf{x}^T)$ com $\mathbf{x}^T = (x_1, ..., x_k)$, tendo-se $p = k + 1$.
O modelo anterior pode ainda ser escrito na foram

Onde:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k = \beta_j x_j + C \qquad (3.27)$$

onde

$$C = \beta_0 + \beta_1 x_1 + ... + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + ... + \beta_k x_k \qquad (3.28)$$

a probabilidade π é dada por:

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)} = g^{-1}(\eta) \qquad (3.29)$$

Com base em (3.27) verifica-se que o parâmetro β_j corresponde à alteração produzida no *logit* pela alteração de uma unidade em x_j desde que:

1. x_j seja uma covariável com efeito linear
2. a interacção de x_j com as restantes covariáveis é nula
3. as restantes covariáveis permaneçam constantes.

Em termos de *odds* tem-se:

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_j x_j + ... + \beta_k x_k) = \exp(\beta_j x_j + C) = \exp(\beta_j x_j)\exp(C) \qquad (3.30)$$

Em termos de *odds ratio* se for considerada uma alteração de *d* unidades em x_j tem-se:

$$\frac{odds\{Y = 1 \mid x_1, ..., x_j + d, ..., x_k\}}{odds\{Y = 1 \mid x_1, ..., x_j, ..., x_k\}} = \frac{\exp(\beta_j (x_j + d))\exp(C)}{\exp(\beta_j x_j)\exp(C)} = \exp(\beta_j d) \qquad (3.31)$$

Ir-se-á concretizar esta interpretação para os parâmetros do modelo de regressão logística considerando apenas alguns casos de forma a não sobrecarregar a exposição.

1. UMA ÚNICA COVARIÁVEL x BINÁRIA

O modelo (3.31) assume a forma:

$$\text{logit}\{Y=1|\,x\}=\log\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1 x \tag{3.32}$$

e tem-se:

$$logit\{Y=1|\,x=0\}=\log\left(\frac{\pi_2}{1-\pi_2}\right)=\beta_0 \tag{3.33}$$

$$logit\{Y=1|\,x=1\}=\log\left(\frac{\pi_1}{1-\pi_1}\right)=\beta_0+\beta_1 \tag{3.34}$$

donde se verifica que β_1 corresponde à diferença dos ***logit*** nos dois conjuntos, isto é,

$$\beta_1=\log OR=\log\left(\frac{\pi_1}{1-\pi_1}\right)-\log\left(\frac{\pi_2}{1-\pi_2}\right) \tag{3.35}$$

2. UMA ÚNICA COVARIÁVEL x CONTÍNUA

Neste caso o modelo assume que o ***log odds*** tem um comportamento linear em função de x.

$$\text{logit}\{Y=1|\,x\}=\log\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1 x \tag{3.36}$$

e um aumento de uma unidade em x corresponde a um aumento do ***odds* ratio** de uma quantidade igual a $\exp(\beta_1)$. Com efeito,

$$\frac{odds\{Y=1|\,x+1\}}{odds\{Y=1|\,x\}}=\frac{\exp(\beta_1(x+1))}{\exp(\beta_1 x)}=\exp(\beta_1) \tag{3.37}$$

3. DUAS COVARIÁVEIS x_1 BINÁRIA E x_2 CONTÍNUA

Consideremos que temos uma covariável dicotómica $x_1\left(x_1=0,x_1=1\right)$correspondente a dois tratamentos e uma covariável contínua x_2. O modelo de regressão logística mais simples é dado por:

$$\text{logit}\left\{Y=1\,|\,\mathbf{x}^T=(x_1,x_2)\right\}=\log\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1x_1+\beta_2x_2 \qquad (3.38)$$

Este modelo assume que:

1. não existe interacção entre o tratamento e a covariável;
2. para cada tratamento a relação entre x_2 e o *log odds* é linear;
3. as rectas têm igual declive.

O *logit* para cada grupo da covariável dicotómica é:

$$\text{logit}\left\{Y=1\,|\,x_1=0,x_2\right\}=\log\left(\frac{\pi_2}{1-\pi_2}\right)=\beta_0+\beta_2x_2 \qquad (3.39)$$

$$\text{logit}\left\{Y=1\,|\,x_1=1,x_2\right\}=\log\left(\frac{\pi_1}{1-\pi_1}\right)=\beta_0+\beta_1+\beta_2x_2 \qquad (3.40)$$

o *odds ratio* é dado por:

$$OR=\frac{\pi_1/\left(1-\pi_1\right)}{\pi_2/\left(1-\pi_2\right)}=\frac{\exp\left(\beta_0+\beta_1+\beta_2x_2\right)}{\exp\left(\beta_0+\beta_2x_2\right)}=\exp(\beta_1) \qquad (3.41)$$

O aumento de uma unidade em x_2 corresponde ao aumento do *odds ratio* de uma quantidade igual a $exp\left(\beta_2\right)$

$$\frac{odds\left\{Y=1\,|\,x_1,x_2+1\right\}}{odds\left\{Y=1\,|\,x_1,x_2\right\}}=\frac{\exp\left(\beta_0+\beta_1x_1+\beta_2\left(x_2+1\right)\right)}{\exp\left(\beta_0+\beta_1x_1+\beta_2x_2\right)}=\exp\left(\beta_2\right) \qquad (3.42)$$

4. UMA COVARIÁVEL DICOTÓMICA $x_1(x_1=0, x_1=1)$ CORRESPODENTE A DOIS TRATAMENTOS E UMA COVARIÁVEL CONTÍNUA x_2 COM A INTERACÇÃO ENTRE O TRATAMENTO E A COVARIÁVEL CONTÍNUA x_2 [3].

O modelo de regressão logística para o caso considerado é dado por:

$$logit\left\{Y=1|\,\mathbf{x}^T=\left(x_1,x_2,x_3\right)\right\}=\log\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3 \qquad (3.43)$$

onde $x_3=x_1\times x_2$.

O modelo com interacção é equivalente a ajustar dois modelos de regressão logística distintos, um para cada grupo, em que a única variável explicativa é x_2.

$$\text{logit}\left\{Y=1|\,x_1=0,x_2\right\}=\log\left(\frac{\pi_2}{1-\pi_2}\right)=\beta_0+\beta_2x_2 \qquad (3.44)$$

$$logit\left\{Y=1|\,x_1=1,x_2\right\}=\log\left(\frac{\pi_1}{1-\pi_1}\right)=\beta_0+\beta_1+\beta_2x_2+\beta_3x_3=\beta_0^{'}+\beta_2^{'}x_2 \qquad (3.45)$$

com $\beta_0^{'}=\beta_0+\beta_1$ e $\beta_2^{'}=\beta_2+\beta_3$

Neste modelo o *odds ratio* é dado por:

$$OR=\frac{\pi_1/\left(1-\pi_1\right)}{\pi_2/\left(1-\pi_2\right)}=\frac{\exp\left(\beta_0+\beta_1+\beta_2x_2+\beta_3x_2\right)}{\exp\left(\beta_0+\beta_2x_2\right)}=\exp(\beta_1+\beta_3x_2) \qquad (3.46)$$

Uma vez que $x_3=x_1\times x_2$ .para $i=2$

5. UMA ÚNICA COVARIÁVEL *x* COM k>2 CATEGORIAS

O modelo de regressão logística é dado por:

$$\log\left(\frac{\pi}{1-\pi}\right)=\beta_0+\sum_{i=1}^{k-1}\beta_iD_i \qquad (3.47)$$

[3] Ver Cabral (2002) Capítulos 9 e 10.

onde D_i (*i=1,...,k-1*) corresponde às *k-1* variáveis indicatrizes definidas, por exemplo, do seguinte modo (a codificação não é única).

$$\left(D_1, D_2, ..., D_{k-1}\right) = \begin{cases} \left(1,0,...0,0\right) & \text{categoria 1} \\ \left(0,1,...,0,0\right) & \text{categoria 2} \\ ... & \\ \left(0,0,...,1,0\right) & \text{categoria k-1} \\ \left(0,0,...,0,0\right) & \text{categoria k} \end{cases} \tag{3.48}$$

Nesta codificação a k-ésima categoria corresponde a que todas as variáveis indicatrizes sejam nulas. À categoria da variável explicativa que se obtém deste modo dá-se o nome de **classe de referência**. A escolha da categoria para a classe de referência é feita de acordo com o modelo em estudo.

β_j (*j=1,...,k-1*) corresponde à diferença dos *logits* entre a classe correspondente à j-ésima categoria e a classe de referência:

$$logit\left\{Y=1 \mid \left(D_1,...,D_j,...,D_{k-1}\right)=\left(0,...,1,...,0\right)\right\} - logit\left\{Y=1 \mid \left(D_1,...,D_{k-1}\right)=\left(0,...,0,...,0\right)\right\} =$$
$$\beta_0 + \beta_j - \beta_0 = \beta_j$$

onde β_0 corresponde ao *logit* da classe de referência.

3.3.2 INFERÊNCIA NO MODELO DE REGRESSÃO LOGÍSTICA

Como o modelo de regressão logística é um modelo linear generalizado o estimador do vector $\boldsymbol{\beta}$ é obtido tal como é descrito na secção 3.2.3 sendo os testes e os intervalos de confiança apresentados na secção 3.2.4 igualmente válidos. No entanto, devido à interpretação dos parâmetros nos diferentes modelos, apresenta-se de seguida os intervalos de confiança para as funções do vector $\boldsymbol{\beta}$. Devido à equivalência entre IC e teste de hipóteses bilaterais estes foram omitidos. Recorde-se que os IC são aproximados.

A validade da aproximação requer que as condições para dados binários sejam verificadas, estas condições podem ser encontradas na página 67 Felgueiras, (2003) [36].

3.3.2.1 Intervalos de Confiança Aproximados

Em determinadas condições de regularidade[4] tem-se que a distribuição assimptótica de $\hat{\beta}_j$, o estimador de máxima verosimilhança de β_j, é:

$$N\left(\beta_j, I_{jj}^{-1}(\boldsymbol{\beta})\right)$$

Designando por $z_{(1-\alpha/2)}$ o quantil (1-α/2) da N(0,1) e $SE\left(\hat{\beta}_j\right)=\sqrt{I_{jj}^{-1}\left(\hat{\boldsymbol{\beta}}\right)}$ tem-se os seguintes intervalos de confiança aproximados ao nível de confiança (1-α)%

Intervalos de confiança aproximados para β_j e para $d\beta_j$

$$\left[\hat{\beta}_j \pm z_{(1-\alpha/2)} SE\left(\hat{\beta}_j\right)\right] \tag{3.49}$$

Intervalos de confiança aproximados para $exp(\beta_j)$ e para $exp(d\beta_j)$

$$\left[\exp\left(\hat{\beta}_j \pm z_{(1-\alpha/2)} SE\left(\hat{\beta}_j\right)\right)\right] \tag{3.50}$$

$$\left[\exp\left(d\hat{\beta}_j \pm z_{(1-\alpha/2)} dSE\left(\hat{\beta}_j\right)\right)\right] \tag{3.51}$$

Intervalo de confiança aproximado para o *logit* $\left(\hat{\eta}_i = g\left(\hat{\pi}_i\right)\right)$:

$$\left[g\left(\hat{\pi}_i\right) \pm z_{(1-\alpha/2)} \sqrt{\mathrm{var}(g(\hat{\pi}_i)}\right] \tag{3.52}$$

Com:

$$var\left(g\left(\hat{\pi}_i\right)\right) = var\left(\hat{\eta}_i\right) = \sum_{j=1}^{p} x_{ij}^2\, var\left(\hat{\beta}_j\right) + \sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} x_{ij} Cov\left(\hat{\beta}_j, \hat{\beta}_k\right) = \mathbf{x}_i^T I^{-1}(\hat{\boldsymbol{\beta}})\mathbf{x}_i$$

[4] Ver página 67 Felgueiras,2003.

$$\operatorname{var}\left(g\left(\hat{\pi}_i\right)\right)=\operatorname{var}\left(\hat{\eta}_i\right)=\sum_{j=1}^{p}x_{ij}^2\operatorname{var}\left(\hat{\beta}_j\right)+\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p}x_{ij}Cov\left(\hat{\beta}_j,\hat{\beta}_k\right)=\mathbf{x}_i^T I^{-1}(\hat{\boldsymbol{\beta}})\mathbf{x}_i \qquad (3.53)$$

Intervalo para π_i

Tendo em atenção que:

$$\pi_i=\frac{\exp\left(\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}\right)}{1+\exp\left(\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}\right)} \qquad (3.54)$$

não é uma função linear de $\boldsymbol{\beta}$ a dedução do IC baseia-se no:

Método-Delta

Seja G_{ij} o vector linha das primeiras derivadas de $\hat{\pi}_i$ em ordem a $\hat{\boldsymbol{\beta}}$. A j-ésima componente do vector é dada por:

$$G_{ij}=\frac{\partial\hat{\pi}_i}{\partial\hat{\beta}_j}=\frac{x_{ij}\exp\left(\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}\right)}{1+\exp\left(\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}\right)},\ j=0,\ldots,k \qquad (3.55)$$

A aproximação da variância de $\hat{\pi}_i$ é $G_{ij}I^{-1}\left(\hat{\boldsymbol{\beta}}\right)G_{ij}^T$ e o intervalo de confiança aproximado $\left((1-\alpha)\times 100\%\right)$ para π_i é dado por:

$$\hat{\pi}_i\pm z_{1-\alpha/2}\sqrt{G_{ij}I^{-1}\left(\hat{\boldsymbol{\beta}}\right)G_{ij}^T} \qquad (3.56)$$

3.3.2.2 Estatísticas para a selecção do modelo

A metodologia descrita na secção 3.2.5 e as estatísticas de teste ou critérios de informação aí indicados são os utilizados quando se procede à selecção de modelos de regressão logística.

A escolha do modelo que melhor se ajusta aos dados requer um processo de construção em várias fases. No caso do modelo de regressão logística existem vários métodos/técnicas de selecção de covariáveis dos quais se vai apresentar os mais utilizados.

3.3.3 TÉCNICAS DE SELECÇÃO DE COVARIÁVEIS

Dos vários métodos de selecção de covariáveis serão aqui abordados o método de "Hosmer e Lemeshow" e o método *stepwise.*

Método de "Hosmer e Lemeshow":

Este método, de acordo com os seus autores, apresenta-se como o mais indicado no caso de existirem muitas covariáveis no estudo como método de triagem das covariáveis.

É formado pelos seguintes passos:

Passo 1: *Criação de modelos de regressão simples para cada uma das covariáveis presentes.*

Neste passo e no caso da covariável em estudo ser do tipo nominal pode-se utilizar uma tabela de contingência ou ajustar um modelo de regressão simple. No caso das covariáveis contínuas apenas a implementação de um modelo de regressão simples será adequado.

Para cada um dos modelos de regressão simples ajustados obtem-se o valor p referente à significância do declive quer através do teste de razão de verosimilhanças quer através do teste de Wald.

Qualquer covariável com valor p <0.25 é uma possível candidata para o modelo múltiplo, assim como qualquer covariável clinicamente relevante. Este valor recomendado poderá parecer muito elevado no entanto irá permitir seleccionar covariáveis que pareçam pouco importantes num modelo de regressão simples, mas que se revelam mais importantes quando passamos a uma regressão múltipla.

Passo 2: *Ajuste do modelo de regressão múltipla.*

Ajusta-se o modelo de regressão múltipla com as covariáveis seleccionadas no passo 1 e avalia-se a significância de cada covariável do seguinte modo:

a) Através da análise do valor p obtido através do teste de Wald para os coeficientes correspondentes às diferentes covariáveis.
b) Através da comparação de cada coeficiente estimado com o coeficiente do modelo que só contém aquela covariável.

As covariáveis que possuem valores p>0.25 ou que apresentam grandes alterações nos coeficientes estimados, devem ser excluídas.

Passo 3: *Ajustamento de um novo modelo múltiplo com as covariáveis seleccionadas no passo 2.*

O novo modelo deve ser comparado com o anterior através do teste de razão de verosimilhanças e novamente avaliada cada covariável incluída no modelo tal como feito no passo 2.

O processo de selecção continua até todas as covariáveis importantes fazerem parte do modelo múltiplo e as excluídas não serem estatisticamente significativas.
Por último deve-se ainda incluir cada covariável que foi excluída no passo 1 por ordem decrescente de valor p, com o objectivo de identificar possíveis covariáveis que por si só não são significativas mas que na presença de outras passam a sê-lo.

Passo 4: *Introdução de interacções.*

Por último devem ser introduzidas no modelo múltiplo todas as interacções que sejam clinicamente relevantes. Uma vez decididas quais as interacções a introduzir no modelo realiza-se o processo de selecção anterior, tendo em atenção que interacções com valores p superiores a 0.10 ou 0.05 e aquelas para as quais haja grandes alterações nos valores dos coeficientes deverão ser excluídas.

Método *stepwise:*

Existem três métodos de selecção *stepwise: forward, backward* e *both*, os três métodos são baseados nos valores p obtidos através de testes de selecção de modelos, em que cada passo se avalia a inclusão ou exclusão de covariáveis ou das suas interacções. Vai-se apresentar uma breve descrição.

Na selecção *forward stepwise*, as variáveis são adicionadas ao modelo, uma de cada vez. Em cada passo a covariável que foi incluída é aquela que proporciona o maior decréscimo no valor da função desvio aquando a sua inclusão. O processo acaba quando a próxima candidata para inclusão não reduz o valor da função desvio até à condição de paragem.

A diferença do *backward stepwise* está no facto de se começar com o modelo saturado e de seguidamente as covariáveis serem excluídas uma de cada vez.

Por último, o *both stepwise* funciona tal como o *forward stepwise*, no entanto uma variável que foi incluída no modelo pode ser excluída num dos passos finais. Desta forma, depois de adicionar uma covariável no modelo o método permite confirmar se algumas covariáveis já incluídas podem ser agora excluídas. Mais uma vez o processo acaba quando a condição de paragem é atingida[35].

3.3.4 VERIFICAÇÃO DA ESCALA DAS COVARIÁVEIS

Uma vez obtido o modelo que parece o mais ajustado aos dados, deve-se analisar cada covariável mais detalhadamente. No caso das covariáveis discretas esta análise deve realizar-se aquando do ajuste dos modelos de regressão simples. Para isto basta verificarmos que o número de observações por categoria seja suficiente de forma a garantir que a frequência de cada célula na tabela de contingência seja ≥ 5.

Para as covariáveis contínuas deve-se verificar a linearidade das mesmas na escala logit. Um método bastante utilizado é o ajustamento dos Modelos Lineares Generalizados Aditivos (MLGA), em que a relação entre a variável resposta e as covariáveis é estimada através da técnica *smoothing*, sendo que depois este ajustamento será comparado com o ajustamento linear[27].

Modelo linear generalizado aditivo:

No caso do MLGA a componente linear, obtida num MLG, é substituída por uma função *smooth*, não paramétrica dada por:

$$g(\mu) = \alpha + \sum_{j=1}^{p} f_i(x_j) \tag{3.57}$$

No caso do modelo logístico tem-se:

$$logit(\pi) = \alpha + \sum_{j=1}^{p} f_i(x_j) \tag{3.58}$$

A estimação da função $f_i(x_j)$ é feita recorrendo às técnicas *smoothing* que permitem que sejam os próprios dados a sugerir a forma de relação entre a variável resposta e as covariáveis.

A estimação da função *smooth* pode ser feita utilizando diferentes métodos de *smoother* ou alisamento: ajustamento local de polinómios, funções Kernel ou ajustamento de regressão robusta pesada, chamado método de *lowess*. O método mais usado consiste em usar *splines*, nomeadamente o *spline* cúbico em que a curva a adoptar é aquela que minimiza para uma covariável x:

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int f''(\mathrm{x})\partial\,\mathrm{x} \qquad (3.59)$$

Onde $f''(\mathrm{x})$ é a segunda derivada de f em ordem a x e λ um parâmetro não negativo de alisamento.

De forma a avaliar a contribuição dos termos não lineares ajustados, o MLGA e o MLG ajustado aos dados podem ser comparados através de um teste F.

3.3.5 MEDIDAS DE QUALIDADE

O *desvio* $D = D(\mathbf{y};\hat{\boldsymbol{\mu}})$ e a *Estatística de Pearson* X^2 são as medidas usadas no MLG e em particular no modelo de regressão logística para avaliar a qualidade do ajustamento do modelo. A *Estatística de Hosmer e Lemeshow* é específica para o modelo de regressão logística.

Desvio

Tal como se viu na secção 3.2.5 o desvio, D, pode ser escrito na forma

$$D = D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\sum_i \omega_i \left\{\left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)\right]\right\} = \sum_i d_i \qquad (3.60)$$

onde d_i mede a diferença dos logaritmos das verosimilhanças observada e ajustada para a i-ésima observação, pelo que é usada como medida de adequabilidade do modelo.

No caso de dados binomias o desvio é dado por[36]

$$D = D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\sum_i \left\{ \left[\tilde{y}_i \log\left(\frac{\tilde{y}_i}{m_i \hat{\pi}_i} \right) + (m_i - \tilde{y}_i) \log\left(\frac{m_i - \tilde{y}_i}{m_i (1-\hat{\pi}_i)} \right) \right] \right\} \qquad (3.61)$$

e a distribuição assimptótica do desvio é a de um qui-quadrado com *n-p* graus de liberdade, embora esta aproximação seja bastante má mesmo para grandes amostras.

Assim, a análise do desvio deve ser considerada como um guia no estudo da adequabilidade do modelo[28] considerando-se que um valor pequeno de D indica um bom ajustamento Esta função tem um papel semelhante aos mínimos quadrados na regressão linear.

Quando $m_i = 1$ tem-se $y_i = \tilde{y}_i$ e a função desvio assume a forma[36].

$$-2\sum_i \left\{ y_i \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i} \right) + \log(1-\hat{\pi}_i) \right\} = -2\left[\hat{\boldsymbol{\beta}}^T \mathbf{x}^T \mathbf{y} + \sum_i \log(1-\hat{\pi}_i) \right] \qquad (3.62)$$

Passando a ser função de $\hat{\boldsymbol{\beta}}$ e como tal não deve ser utilizada como medidade adequabilidade.

Estatística de Pearson

A outra medida de adequabilidade é a estatística de Perason generalizada (3.2.3) dada por:

$$X^2 = \sum_i \frac{\omega_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \qquad (3.63)$$

que para dados binomias se escreve na forma:

$$X^2 = \sum_{i=1}^{g} \frac{(\tilde{y}_i - m_i \hat{\pi}_i)^2}{m_i \hat{\pi}_i (1-\hat{\pi}_i)} \qquad (3.64)$$

onde g é o número de grupos (classes covariáveis). A adequabilidade do modelo é testada comparando X^2 com o quantil de probabilidade $1-\alpha$ de uma distribuição qui-quadrado com $g-p-1$ graus de liberdade. Esta aproximação pode ser de má qualidade mesmo para grandes amostras[36].

Quando $m_i = 1$ então $\hat{\pi} = \bar{y}$ e:

$$X^2 = \sum_i \frac{(y_i - \bar{y})^2}{\bar{y}(1-\bar{y})} = n \qquad (3.65)$$

e portanto não faz sentido usar esta estaística neste caso.

Estatística de Hosmer e Lemeshow

A vantagem desta medida surge do facto de poder ser aplicada a dados binários, mais especificamente para dados não agrupados.

Para calcular esta estatística as observações são ordenadas por ordem crescente do valor obtido para a probabilidade ajustada pelo modelo e posteriormente agrupadas.

Dois métodos de criar grupos foram propostos:
(1) Baseado nos percentis das probabilidades ajustadas;
(2) Baseado num número fixo de probabilidades ajustadas.

No primeiro método (1) se os dados forem agrupados em 10 grupos, $g = 10$, o primeiro grupo será constituído por $n_1^* = n/10$ indivíduos, que possuem os menores valor para as probabilidades ajustadas e o último grupo será constituído então por $n_{10}^* = n/10$ indivíduos que possuem os valores mais elevados para as probabilidades ajustadas.
No segundo método (2) a utilização de $g = 10$ resulta em pontos de corte definidos para os valores $k/10, k = 1, 2, ..., 9$, e os grupos contêm todos os indivíduos com probabilidades ajustadas entre os pontos de corte.

Qualquer que seja a forma de agrupar os dados a estatística é dada por:

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n_k^* \bar{\pi}_k)^2}{n_k^* \bar{\pi}_k (1-\bar{\pi}_k)} \sim \chi^2_{g-1} \qquad (3.66)$$

onde n_k^* é o número de indivíduos no k-ésimo grupo, O_k é o número de sucessos observados, e $\hat{\pi}_k$ é dada por: $\hat{\pi}_k = e_k / n_k^*$ em que e_k é o número de sucessos esperado estimado.

No entanto, uma vez que o valor obtido para a estatística de teste depende do número de grupos escolhidos, do número de observações por cada grupo e de como o conjunto de

observações com valores de probabilidades ajustadas são divididas, o valor p obtido não deve ser interpretado de forma rigorosa mas sim encarado como uma medida informal de avaliação do ajustamento do modelo[35].

3.3.6 Análise de Resíduos

Uma vez ajustado o modelo logístico aos valores observados de uma variável resposta binária ou binomial é essencial verificar se é realmente válido no que diz respeito à escolha da distribuição, da função de ligação e em termos do preditor linear, assim como na identificação de observações mal ajustadas. Essa análise é feita com base nos resíduos.

Resíduos

São medidas de concordância entre o valor observado para a variável resposta e o valor ajustado.

Tendo em conta um vector com os valores observados para a variável resposta $\tilde{\mathbf{y}} = (\tilde{y}_1, ..., \tilde{y}_n)^T$, e um outro vector com os valores ajustados pelo modelo $\hat{\tilde{\mathbf{y}}} = \left(\hat{\tilde{y}}_1, ..., \hat{\tilde{y}}_n\right)^T$, o vector dos resíduos será então obtido pela diferença de:

$$\hat{r} = (\tilde{y}_1 - \hat{\tilde{y}}_1, ..., \tilde{y}_n - \hat{\tilde{y}}_n)^T$$

onde $\hat{\tilde{y}}_i = m_i \hat{\pi}_i$.

Diferentes tipos de resíduos podem ser obtidos apresentando-se aqui somente a descrição daqueles que são utilizados na análise dos dados apresentada no Capítulo 4 e que podem ser utilizados quer os dados sejam agrupados ou não ($m_i = 1 \text{e } y_i = \tilde{y}_i$)

Uma descrição mais detalhada sobre resíduos poderá ser consultada[28, 35-36] .

Resíduos de Pearson: Estes resíduos são obtidos da seguinte forma:

$$X_i = \frac{\tilde{y}_i - m_i \hat{\pi}_i}{\sqrt{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}} \qquad (3.67)$$

são conhecidos como resíduos de Pearson uma vez que $X^2 = \sum_i X_i^2$ sendo esta a estatística de Pearson. Os resíduos assim obtidos podem não ter variância 1. Atendendo a que $var\left(Y_i - \hat{Y}_i\right) \approx var(Y_i)(1-h_i)$, onde h_i é o *leverage*[5]. O resíduo de Pearson padronizado é dado por:

$$r_{p_i} = \frac{\tilde{y}_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i\left(1-\hat{\pi}_i\right)(1-h_i)}} \tag{3.68}$$

Resíduos Desvio: São construídos a partir da função desvio obtida depois de ajustar o modelo de regressão logística. O desvio residual é dado por:

$$D_i = \text{sgn}(\tilde{y}_i - \hat{\tilde{y}}_i)\left\{2y_i \log\left(\frac{\tilde{y}_i}{\hat{\tilde{y}}_i}\right) + 2(m_i - \tilde{y}_i)\log\left(\frac{m_i - \tilde{y}_i}{m_i - \hat{\tilde{y}}_i}\right)\right\}^{1/2} \tag{3.69}$$

onde $\text{sgn}(\tilde{y}_i - \hat{\tilde{y}}_i)$: designa o sinal de $\tilde{y}_i - \hat{\tilde{y}}_i$.

A padronização nos resíduos de desvio é também feita com recurso ao l*everage* e é dada por:

$$r_{D_i} = \frac{D_i}{\sqrt{(1-h_i)}} \tag{3.70}$$

Resíduos de Verosimilhança: Estes resíduos são obtidos através da comparação entre o desvio verificado quando se ajusta o modelo às n observações, e o desvio obtido quando se aplica o mesmo modelo a $n-1$ observações, isto é excluindo o i-ésimo elemento da amostra.

Este procedimento é computacionalmente muito exigente pelo que é usado uma aproximação dada por $h_i r_{p_i}^2 + (1-h_i)r_{D_i}^2$. A partir desta quantidade e da combinação entre os resíduos de Pearson e os resíduos desvio obtem-se o resíduo de verosimilhança:

$$r_{L_i} = \text{sgn}(\tilde{y}_i - \hat{\tilde{y}}_i)\sqrt{h_i r_{p_i}^2 + (1-h_i)r_{D_i}^2} \tag{3.71}$$

[5] O *Leverage* h_i procura medir quão extremos são os valores das covariáveis para o *i*-ésimo indivíduo, em relação aos restantes.

3.2.5.1 Análise gráfica dos resíduos

Uma vez ajustado um modelo de regressão pode verificar-se a existência de "anomalias", tanto na componente aleatória como na sistemática, estas podem ser detectadas através da análise de representações gráficas dos resíduos

Avaliação do preditor linear e identificação de *outliers*

Quando existe adequação do preditor linear os gráficos dos resíduos padronizados *versus* o preditor linear, os valores ajustados ou o índice do resíduo devem mostrar um padrão nulo, em que os resíduos estão distribuídos em torno do zero com amplitude constante para diferentes valores de $\hat{y}_i$.

Diferentes tipos de gráficos podem ser aplicados, são aqui referidos os mais usados para a regressão logística.

Gráfico dos resíduos vs o índice

É considerado o gráfico mais simples e corresponde à representação dos resíduos contra o seu número de observação. Apesar de este gráfico ter maior utilidade na detecção de *outliers*, uma vez que permite identificar os resíduos com valores anormalmente elevados, pode ser também usado para avaliar o ajustamento do modelo. Se o gráfico apresentar um padrão sistemático é indicação de que o modelo não é correcto.

Gráfico das probabilidades cruzadas

Este gráfico é indicado para a detecção de *outliers* no caso em que os dados são binários e consiste no cálculo da probabilidade da i-ésima observação sabendo as restantes $n-1$ observações. A esta probabilidade chama-se probabilidade cruzada e designa-se por: $P\left(\tilde{y}_i \mid \tilde{\mathbf{y}}_{(i)}\right)$ onde $\tilde{\mathbf{y}}_{(i)}$ representa o conjunto de dados excluindo a i-ésima observação. Esta probabilidade pode ser aproximada por:

$$P\left(\tilde{y}_i \mid \tilde{\mathbf{y}}_{(i)}\right) \approx \exp\left(\frac{-r_{L_i}^2}{2}\right)\sqrt{1-h_i} \qquad (3.72)$$

Este gráfico indica as observações com probabilidade reduzida de ocorrerem, sendo por isso consideradas como possíveis *outliers*.

Gráfico *half-normal* dos resíduos:

Os resíduos r_{L_i} e r_{D_i} seguem uma distribuição aproximadamente normal padrão, quando o modelo ajustado é o correcto pelo que o papel de probabilidades normal poderá ser útil na avaliação do modelo.

Embora este gráfico tenha sido pensado para avaliar o pressuposto de normalidade dos resíduos, nos modelos de regressão logística é mais útil para avaliar se o modelo em análise é ou não adequado e revelar a presença de *outliers*.

A utilização do papel de probabilidades *half-normal* mostra, de forma mais efectiva, estes aspectos.

Deve-se no entanto ter em atenção que mesmo quando o modelo ajustado é correcto os resíduos utilizados na construção do papel de probabilidades *half-normal* são correlacionados e podem não ter uma distribuição aproximada normal. Por isso a sua representação não será necessariamente sobre uma linha recta.

Na interpretação deste gráfico, tem todo o interesse a construção de um envelope simulado. Este envelope é tal, que se o modelo ajustado é o correcto, existe grande probabilidade de que todos os pontos do gráfico estejam contidos dentro dos limites. Este aspecto é útil na detecção de *outliers* que, no caso de existirem, aparecerão no canto superior direito do gráfico, separados dos restantes. No entanto, a principal vantagem do envelope é o facto do papel de probabilidades *half-normal* poder ser interpretado sem ser feita qualquer hipótese sobre a distribuição dos resíduos.

A descrição da construção do envelope pode ser vista em Collett[35] na página 129.

Avaliação da função de ligação:

Um método para verificar se a função de ligação escolhida é adequada consiste em considerar $\hat{\eta}^2$ como uma nova covariável a adicionar ao preditor linear e verificar se há declínio na função desvio e se este é significativo. Este método foi sugerido por Hinkley (1995)[28].

Identificação de observações influentes:

Uma observação é influente se a sua exclusão produz alterações significativas nas estimativas dos parâmetros do modelo. A sua presença pode, por isso, originar um impacto indevido nas conclusões retiradas do modelo.

Quando uma observação é distante das restantes observações em termos das covariáveis explicativas pode ser considerada uma observação influente. Esta distância entre a i-ésima observação em relação às restantes observações é medida geralmente pelo *leverage* h_i.
Algumas representações gráficas úteis na identificação de observações influentes são as seguintes:

Gráfico do *leverage* em função do seu número de ordem:

Consiste em representar os h_i *versus* o seu número de ordem. Uma observação é considerada influente se $h_i > 2p/n$, em que p é o número de parâmetros no modelo.

Gráfico da estatística D:

De forma a avaliar a influência da i-ésima observação no vector $\mathbf{b}$ das estimativas dos parâmetros, obtido com base em todas as observações, ajusta-se ao modelo ao conjunto de dados ao qual se retirou essa observação, obtendo-se assim uma nova estimativa, $\mathbf{b}_{(i)}$ daquele vector. A comparação entre $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\beta}}_{(i)}$ é feita com base na estatística:

$$D_i = \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(i)})^T \mathbf{X}^T \mathbf{W} \mathbf{X} (\hat{\beta} - \hat{\beta}_{(i)}) \qquad (3.73)$$

Em que $\mathbf{X}$ é a matriz das covariáveis, $\mathbf{W}$ é a matriz de pesos (3.15) e p é o número de parâmetros no modelo. Esta estatística dá o quadrado da distância entre aqueles estimadores

O cálculo de $\mathbf{b}_{(i)}$ pode ser aproximado por:

$$\mathbf{b}_{(i)} \approx \mathbf{b} - (1-h_i)^{1/2} \hat{\omega}_i^{1/2} r_{pi} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{x}_i \qquad (3.74)$$

Tendo em conta esta expressão podemos descrever D_i da seguinte forma:

$$D_i = \frac{h_i r_{p_i}^2}{p(1-h_i)} \qquad (3.75)$$

Segundo Collett (2003) esta estatística é a mais útil para avaliar a influência de uma observação na estimativa dos parâmetros do modelo. Valores relativamente elevados para D_i indicam, que a respectiva observação é influente.

3.2.6 AVALIAÇÃO DA CAPACIDADE PREDITIVA DO MODELO

A capacidade preditiva de um modelo consiste na sua capacidade de discriminar entre os indivíduos que de facto experimentaram o acontecimento de interesse daqueles que não o experimentaram. No entanto é de referir que um modelo que se ajusta bem aos dados não tem necessariamente uma boa capacidade preditiva.

Hosmer e Lemeshow sugerem algumas técnicas usadas para a avaliação da capacidade preditiva do modelo, entre elas:

Erro de predição

Consiste em definir um ponto de corte, sendo 0.5 o mais comum, e desta forma criar uma variável dicotómica em que os valores ajustados são agrupados do seguinte modo:

Se os valores ajustados forem superior a 0.5 $\Rightarrow$ a nova variável assumirá o valor 1.

Se os valores ajustados forem inferior a 0.5 $\Rightarrow$ a nova variável assumirá o valor 0.

Esta nova variável é cruzada com os valores observados permitindo assim calcular a proporção de casos preditos correctamente.

Curva ROC

Este gráfico permite detectar os verdadeiros positivos, indivíduos com teste positivo nos quais existe realmente o sintoma, e verdadeiros negativos, dados por 1-especificidade, sendo que a especificidade é definida como a probabilidade do teste dar negativo dado que o indivíduo não tem sintomas, para diferentes pontos de corte[34].

A área sob a curva (AUC do inglês *Area under the curve*) ROC, que varia entre 0 e 1, classifica de forma mais precisa a capacidade de discriminação do modelo, isto é, fornece uma medida de discriminação que consiste na capacidade de distinguir um indivíduo em que a resposta de interesse se verifica dos indivíduos em que tal não acontece. A Tabela 1 mostra as diferentes classificações para os valores obtidos para a AUC:

Tabela 1: Classificação para os valores obtidos para a AUC[34].

AUC	Diagnóstico
AUC =0.5	Modelo sem poder discriminativo
0.7≤AUC<0.8	Discriminação aceitável
0.8≤AUC<0.9	Discriminação excelente
AUC≥0.9	Discriminação extraordinária

Capítulo 4

Modelação do tempo total de isquémia

4.1 OS DADOS

Os dados do presente estudo foram recolhidos no Serviço de Cardiologia I do Hospital de Santa Maria (HSM) e teve como população alvo os doentes com diagnóstico confirmado de EAMEST, que recorreram ao Serviço de Urgência (SU) do HSM, nos quais foi restabelecido o fluxo por ICP primária, entre o dia 1 de Janeiro de 2010 a dia 31 de Dezembro de 2010.

O protocolo deste estudo foi apresentado à Comissão de Ética do HSM, no dia 30 de Junho de 2011 e obteve parecer favorável no dia 25 de Outubro de 2011.

A recolha de dados foi efectuada através da consulta dos processos clínicos, que se encontravam no Serviço de Cardiologia I, estes dados foram registados num primeiro questionário (Apêndice 1). Foi também aplicado um segundo questionário efectuado por via telefónica, ao doente ou seu cuidador (Apêndice 2).

Foram estabelecidos critérios de inclusão e de exclusão, para que desta forma fosse possível obter uma amostra homogénea e representativa. Apenas os doentes em que todos os dados do primeiro questionário estivessem completos foram contactados por via telefónica, com o intuito de adquirir os dados do segundo questionário.

Critérios de inclusão:

- Diagnóstico confirmado de SCA, mais especificamente EAMEST;
- Indivíduos que tenham realizado ICP primária.

Critérios de exclusão:

- Indivíduos que se recusassem a participar no estudo;
- Indivíduos cujo o primeiro questionário não estivesse completo;
- Indivíduos que apresentassem incapacidade física ou cognitiva para responder ao segundo questionário;
- Doentes em que o número de telefone não fosse válido.

A variável resposta (Y) foi definida como o tempo total de isquémia, que assume dois valores: Y=1 se o tempo total de isquémia é superior a 6 horas, também dito tempo total aumentado, e Y=0 no caso contrário. O objectivo deste trabalho é o de modelar o tempo total de isquémia querendo com isto dizer-se, identificar quais as covariáveis que contribuem para a ocorrência de um tempo total de isquémia aumentado.

Toda a informação foi tratada de forma confidencial, e os dados foram introduzidos numa base de dados, com identificação codificada, de forma a poderem ser efectuadas análises estatísticas.

Questionários:

O primeiro questionário é constituído por 15 perguntas, e diz respeito aos dados obtidos através da consulta dos processos clínicos, nomeadamente dados demográficos, idade, sexo e raça. Dados sobre a presença de: factores de risco, antecedentes de SCA, zona onde se encontrava no momento de início da sintomatologia, o facto de ter sido transferido de outro hospital e ainda o registo do dia da semana da ocorrência do EAMEST. As horas do início da sintomatologia e início do tratamento foram também registadas de forma a permitir o cálculo do tempo total de isquémia.

O segundo questionário diz respeito aos dados obtidos por telefone, constituído por 9 itens. Trata-se de questões relacionadas com o nível socioeconómico, através da aplicação da escala de Graffar (Anexo 1), que é um instrumento de caracterização demográfica constituído por cinco domínios que caracterizam o nível socioeconómico-cultural de cada indivíduo: profissão, grau de instrução, origem dos rendimentos, qualidade da habitação e tipo de zona residencial. Em cada domínio, são apresentadas cinco categorias de resposta (previamente definidas), sendo atribuído, a cada uma, um valor de 1 a 5. A pontuação total varia, assim, entre 5 e 25, sendo dividida em cinco intervalos, correspondendo cada um a uma classe ou nível social: por 5 níveis, ao nível I corresponde o nível socioeconómico alto; ao nível II, o nível socioeconómico médio alto; ao III, o nível socioeconómico médio; ao IV, o nível socioeconómico médio baixo e ao V, o nível socioeconómico baixo.

No segundo questionário também foi registado o nível de dor experimentado por cada indivíduo, através da aplicação da escala numérica da intensidade da dor (Anexo 2). Esta escala consiste numa régua dividida em onze partes iguais, numeradas sucessivamente de 0 a 10.

Pretende-se que o doente faça a equivalência entre a intensidade da sua dor e uma classificação numérica, sendo que 0 corresponde a classificação “Sem Dor” e 10 a classificação “Dor Máxima” (Dor de intensidade máxima imaginável)[37].

Obteve-se também variáveis como o nível de escolaridade, se possuía conhecimento sobre a doença, se estava acompanhado no início da sintomatologia e as actividades que estava a desempenhar.

4.2 CODIFICAÇÃO DOS DADOS

Todas as variáveis obtidas, com excepção da idade, foram categorizadas tendo em conta o seu significado clínico e no que diz respeito ao tempo total de isquémia seguindo as recomendações da American Heart Association (AHA).

De seguida apresenta-se a Tabela 2 com as diferentes variáveis e a sua respectiva categorização, assim como a frequência de indivíduos por categoria. Na última coluna está representado, de acordo com as categorias, o número de indivíduos cujo tempo total de isquémia é superior a 6 horas, por tratar-se do acontecimento de interesse para a variável resposta.

Tabela 2: Categorização das variáveis em estudo e frequência de indivíduos por categoria.

Nome da variável	Codificação	Frequência de indivíduos	Número de indivíduos com tempo total >6 (n=54)
Tempo total	**0** - < 6 horas	74	Não se aplica.
	1 - ≥ 6 horas	54	
Idade (anos)*	0 - < 55	42	10
	1 - entre 55-75	64	30
	2 - >75	22	14
Sexo	**0** - sexo feminino	33	14
	1 - sexo masculino	95	40
Raça	**0** - Caucasiana	123	51
	1 - Outras	5	3
Nível socioeconómico	**1** - Alto	13	3
	2 - Médio alto	28	7
	3 - Médio	33	12
	4 - Médio baixo	45	25
	5 - Baixo	9	7
Nível de escolaridade	**1** - Ensino primário ou sem escolaridade	73	39
	2 - Ensino secundário	27	9
	3 - Ensino Superior	28	6
Conhecimento da doença	**0** – Não	61	33
	1 – Sim	67	21
Antecedentes de EAMEST	**0** – Não	106	44
	1 – Sim	22	10
Factores de Risco para EAMEST	**0** – Não	2	1
	1 – Sim	126	53
Escala da dor	**3** - Nível 3 de dor**	8	6
	4 - Nível 4 de dor	20	14
	5 - Nível 5 de dor	8	3
	6 - Nível 6 de dor	28	15
	7 - Nível 7 de dor	14	4
	8 - Nível 8 de dor	22	7
	9 - Nível 9 de dor	14	7

	10 - Nível 10 de dor	12	3
Tipo de companhia na altura do EAMEST	**0** - Familiares	67	37
	1 - Amigos	18	3
	2 - Colegas de trabalho	9	1
	3 - Não acompanhado	34	13
Funções que desempenhava na altura do EAMEST	**0 -** Lazer	18	4
	1 - Trabalhar	20	7
	2 - Esforço físico	20	6
	3 - Dormir	70	37
Zona em que se encontrava	**0** - Urbana	115	45
	1 - Rural	13	9
Transferido de outro hospital	**0** - Não	109	43
	1 - Sim	19	11
Dia de ocorrência do EAMEST	**0** - Dia de semana	87	39
	1 - Fim de semana ou feriado	41	15

*Note-se que a covariável idade apesar de ter sido obtida na sua forma contínua foi categorizada, de acordo com a literatura publicada, de forma a facilitar à sua interpretação clínica[38-40]. Ver Apêndice 3 para mais detalhes.

**Para a variável Escala da dor, só se registaram níveis de dor a partir do nível 3 de dor, isto é, não existiram registos de níveis de dor correspondentes aos níveis, 0, 1 ou 2.

É importante referir que o conjunto de dados aqui estudado, encontra-se de forma não agrupada, pois existe um conjunto diferente de covariáveis para cada indivíduo.

4.3 CARACTERIZAÇÃO DA AMOSTRA

A amostra é constituída por 128 doentes, 74.22% dos quais do sexo masculino e maioritariamente de raça caucasiana, com 123 indivíduos nesta categoria. A idade média foi de aproximadamente 62 anos (61.7). Mais de metade dos indivíduos, 57.03%, tinha apenas o ensino primário ou eram analfabetos (Figura 2), e por último, a maioria dos sujeitos eram de classe social média baixa, 35.16% (Figura 3).

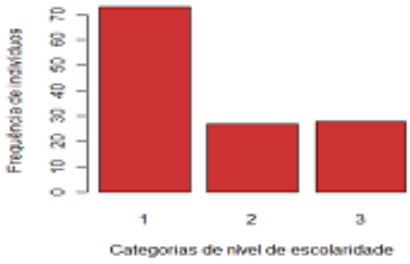

Figura 1: Distribuição dos indivíduos pelas categorias do nível de escolaridade.

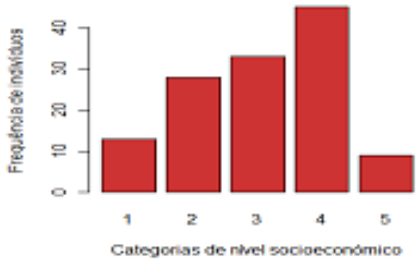

Figura 2: Distribuição dos indivíduos pelas categorias do nível socioeconómico.

Dos 128 doentes estudados, 54 tinham um tempo total superior a 6 horas, que correspondem a 42.19% da amostra.

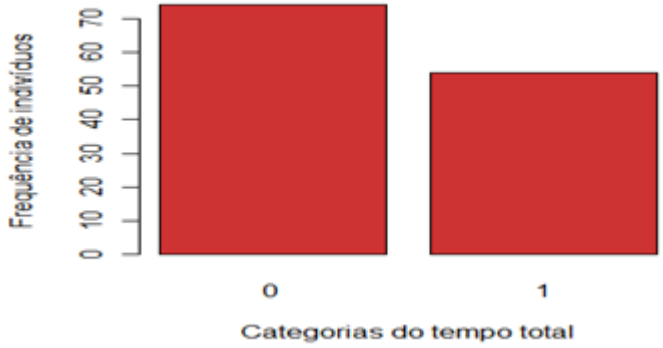

Figura 2: Distribuição dos indivíduos pelas categorias do tempo total.

Verificou-se que apenas 2 doentes não possuíam nenhum tipo de factor de risco para SCA e ainda que 106 doentes não tinham antecedentes de SCA.

4.4 SELECÇÃO DE COVARIÁVEIS

Previamente ao ajuste do modelo foi realizada, para as covariáveis categóricas, uma tabela de contingência, uma vez que como recomendado por Hosmer e Lemshow[34] deve ter-se em especial atenção as células da tabela que apresentem valores nulos, já que irá condicionar uma estimativa pontual dos OR a tender ou para zero ou para

infinito. Deve-se também, ter atenção que apenas 20% dos valores esperados para cada célula podem ser inferiores a 5[41].

Após a avaliação cuidadosa das tabelas de contigência chegou-se à conclusão que as covariáveis **raça** e **factores de risco** deviam ser retiradas da análise, por terem valores esperados inferiores a 5 em mais de 20% das células e possuirem somenente 2 categorias o que impossibilita o seu reagrupamento.

Para a covariável **nível socioecónomico**, foi também verificada a existência de valores esperados inferiores a 5 em mais de 20% das células. Neste caso optou-se por reagrupar duas das categorias, nível 4 e 5 da covariável, numa única categoria que passará a ser designada por nível baixo (Tabela 3).

Tabela 3: Categorias da covariável Nível socioeconómico após ter sido reagrupada.

	Codificação	Frequência de indivíduos	Número de indivíduos com tempo total >6h
Nível socioeconómico	**1** - Superior	13	3
	2 - Médio alto	28	7
	3 - Médio	33	12
	4 - Baixo	54	32

Depois de uma análise a cada covariável obteve-se então as possíveis candidatas a variáveis explicativas no nosso modelo, que passamos a descrever:

- Idade
- Sexo
- Nível socioeconómico
- Nível de escolaridade
- Conhecimento da doença
- Factores de risco
- Zona
- Transferido
- Dia
- Escala da dor
- Antecedentes pessoais.

Definida a variável resposta e as possíveis variáveis explicativas passou-se à construção do modelo.

Os métodos de selecção de covariáveis aplicados ao conjunto de dados foram: o método proposto por "Hosmer e Lemeshow" e os métodos *stepwise.*

Aplicação dos Métodos de selecção de covariáveis

4.4.1 Método de "Hosmer e Lemeshow"

Passo 1: *Criação de modelos de regressão simples para cada uma das covariáveis presentes.*

Foi realizado o ajuste de modelos univariados para cada uma das covariáveis. Para cada modelo simples ajustado é apresentado a estimativa dos coeficientes, o valor para o teste Wald e para o Teste de Razão de Verosimilhança e os respectivos valores p (Tabela 4).

Tabela 4: Modelos simples ajustados para cada uma das covariáveis

Covariável	Estimativa Coeficiente	Teste de Wald	valor p (Wald)	*Deviance*	valor p (Razão de Verossimilhança)
Idade 1	1.038	9.9135	0.007036	10.889	0.004321
Idade 2	1.7228				
Sexo 1	-0.01307	0.001	0.9745	0.0010214	0.9745
Dia 1	-0.3424	0.7734	0.3792	0.78255	0.3764
Transferido 1	0.7469	2.1936	0.1386	2.2277	0.1356
Zona 1	1.2528	3.947	0.04695	4.3141	0.0378
Antecedentes 1	0.1606	0.1161	0.7333	0.11569	0.7338
Nível socioeconómico 2	0.1054				
Nível socioeconómico 3	0.6444	11.673	0.00859	12.513	0.005818
Nível socioeconómico 4	1.5787				
Nível escolaridade 2	-0.8303	9.0806	0.01067	9.9828	0.006796
Nível escolaridade 3	-1.4365				
Tipo de companhia 1	-1.8192				
Tipo de companhia 2	-2.2892	11.5	0.009308	14.426	0.002379
Tipo de companhia 3	-0.6893				
Funções 1	0.6337				
Funções 2	0.4055	7.5319	0.05675	8.0941	0.04411
Funções 3	1.3672				
Conhecimento da doença 1	-0.9484	6.6445	0.009946	6.8303	0.008962
Escala da dor 4	-0.2513				
Escala da dor 5	-1.6094				
Escala da dor 6	-0.9555				
Escala da dor 7	-2.0149	18.166	0.01124	21.313	0.003334
Escala da dor 8	-1.8608				
Escala da dor 9	-2.8904				
Escala da dor 10	-2.3979				

Como sugerido pelo Hosmer e Lemeshow, foram seleccionadas as variáveis com valor p inferior a 0.25. Após a aplicação deste critério obtiveram-se as seguintes covariáveis:

- Idade
- Transferido
- Zona
- Nível socioeconómico
- Nível escolaridade
- Tipo de companhia
- Funções
- Conhecimento da doença
- Escala da dor.

Passo 2: *Ajuste do modelo de regressão múltipla.*

Identificadas as possíveis variáveis explicativas passou-se à construção do modelo múltiplo (Model 2) contendo todas as variáveis seleccionadas. Na Tabela 5 são apresentadas as estimativas para os coeficientes, o valor para os testes Wald e os valores p resultantes destes.

Tabela 5: Modelo múltiplo (Model 2) ajustado após selecção das covariáveis a partir dos modelos de regressão simples.

Covariável	Estimativa Coeficiente	Teste de Wald	Valor p (Wald)	*Deviance*	valor p (Razão de Verosimilhanças)
Idade 1	1.4234	5.7	0.057	6.1318	0.04661
Idade 2	1.6581				
Transferido 1	-1.6438	3	0.083	3.2043	0.07344
Zona 1	2.3925	4.4	0.036	4.7574	0.02917
Nível socioeconómico 2	-0.1998	5.4	0.15	5.807	0.1214
Nível socioeconómico 3	-0.4543				
Nível socioeconómico 4	1.1193				
Nível escolaridade 2	0.653	1.1	0.57	1.1428	0.5647
Nível escolaridade 3	-0.2826				
Tipo de companhia 1	-1.929	4.2	0.24	4.6319	0.2008
Tipo de companhia 2	-2.0189				
Tipo de companhia 3	-0.3287				
Funções 1	1.0176	6.9	0.075	7.7047	0.05253
Funções 2	-1.0642				
Funções 3	0.5739				
Conhecimento da doença 1	-0.3544	0.4	0.53	0.40039	0.5269
Escalador da 4	-1.2482	15.1	0.035	18.859	0.008639
Escalador da 5	-2.0012				
Escalador da 6	-1.607				
Escala da dor 7	-2.3062				
Escala da dor 8	-3.0379				
Escala da dor 9	-3.5288				
Escala da dor10	-3.8825				

De seguida, e para cada covariável, analisou-se o valor p obtido com base no teste de Wald e as estimativas dos coeficientes que foram comparadas com as obtidas no passo 1 através dos modelos de regressão simples. As covariáveis com valores p <0.25 ou que

não apresentaram grandes alterações nos coeficientes estimados, foram as seleccionadas, a saber:

- Idade,
- Transferido,
- Zona,
- Funções,
- Nível socioeconómico,
- Tipo de companhia,
- Escala da dor.

Passo 3: *Ajustamento de um novo modelo múltiplo com as covariáveis seleccionadas no passo 2.*

Um novo modelo múltiplo (Model 1) foi ajustado com aquelas covariáveis. Os resultados obtidos são apresentados na Tabela 6.

Tabela 6: Modelo múltiplo (Model 1) ajustado após a selecção das covariáveis.

Covariável	Estimativa Coeficiente	Teste de Wald	Valor p (Wald)	*Deviance*	Valor p (Razão de verosimilhanças)
idade 1	1.0827	5	0.081	5.3756	0.06803
idade 2	1.3087				
Zona 1	1.0461	3.9	0.049	4.1803	0.0409
Funções 1	1.0188				
Funções 2	-1.1125	7	0.072	7.8297	0.04967
Funções 3	0.4452				
Transferido 1	-1.5665	2.8	0.097	2.9081	0.08814
Nível socioeconómico2	0.1464				
Nível socioeconómico3	0.2699	8.7	0.033	9.3977	0.02445
Nível socioeconómico4	1.5071				
Tipo companhia1	-1.8299				
Tipo companhia2	-2.1614	4.4	0.22	4.8719	0.1814
Tipo companhia3	-0.299				
Escala da dor 4	-1.358				
Escalada dor 5	-2.6181				
Escala da dor 6	-1.4757				
Escala da dor 7	-2.4342	16	0.025	20.183	0.005188
Escala da dor 8	-2.8251				
Escala da dor 9	-3.2826				
Escala da dor 10	-3.5492				

Os modelos (Model 1 e Model 2) foram comparados através do Teste de Razão de Verosimilhanças apresentado-se o respectivo *output*:

```
Model 1: tempototal ~ idadecat2 + zona + funções + transferido +
nívelsócioeconómico + tipocompanhia + escalador
Model 2: tempototal ~ tipocompanhia + escalador + idadecat2 +
nívelsócioeconómico + nivelescolaridade2 + conhecimentodoença + zona +
funções + transferido
        Resid. Df Resid. Dev Df Deviance P(>|Chi|)
      1       107     117.11
      2       104     115.65  3   1.4577    0.6921
```

Nota: Model 1= modelo 1; Model 2= modelo 2.

O modelo 1 foi selecionado com base no valor p do Teste de Razão de Verosimilhanças. Model 1: Tempo total ~ idadecat + transferido + zona + funções + nível socioeconómico + tipo de companhia + escala da dor.

De acordo com o método de "Hosmer e Lemeshow", descrito na secção 3.3.3, mantém-se todas as covariáveis (Model 1). Seguindo o procedimento indicado no referido método, foi introduzido no modelo (Model 1) cada uma das variáveis que tinham sido dispensadas na análise univariada, uma vez que podem passar a ser significativas quando controladas por outras covariáveis. Os resultados não são aqui apresentados, uma vez que nenhuma das covariáveis acrescentadas no modelo revelou ser significativa, nem causou discrepâncias nos coeficientes.
Depois de obtido um modelo múltiplo o passo seguinte foi a introdução de interacções.

Passo 4: *Introdução de interacções:*

No presente caso apenas uma interacção foi introduzida no modelo, a interacção entre as diferentes classes etárias e a escala da dor. A introdução de apenas esta interacção deve-se ao facto de a literatura da especialidade apresentar a possibilidade de existir uma diminuição da sensação da dor com o aumento da idade.
Quando a interacção entre a escala da dor e a variável idade foi introduzida no modelo verificou-se a existência de valores nulos em algumas das categorias de intersecção entre as duas covariáveis[6]. Tendo em conta que, células com valores nulos poderão levar a uma estimativa pontual dos OR a tender para zero ou para infinito, optou-se por retirar a interacção, obtendo-se como modelo final:

Tempo total ~ idadecat + transferido + zona + funções + nível socioeconómico + tipo de companhia + escala da dor.

4.4.2 Método *stepwise*

O modelo logístico inicial incluiu todas as covariáveis. A interacção entre a covariável idade, na sua forma categórica, e a escala da dor foi omitida pelos motivos anteriormente explicados.

Modelo inicial:

[6] Ver apêndice 3

Tempo total ~ idadecat+escala da dor + sexo + raça + dia + transferido + zona + antecedentes + factores de risco + nível socioeconómico+ nível escolaridade + tipo companhia + funções + conhecimento da doença.

Seguiu-se a aplicação dos 3 métodos de selecção, começando pelo *both stepwise*, de seguida o método *backward stepwise* e por último o *forward stepwise.*

Método *both stepwise*

Os resultados obtidos pelo primeiro método aplicado, *both stepwise*, encontram-se na Tabela 7.

Tabela 7: Modelo obtido a partir da aplicação do método *both stepwise.*

Covariável	Estimativa coeficiente	Teste de Wald	Valor p (Wald)	*Deviance*	Valor p (Razão de Verosimilhança)
idade 1	1.2883	5.6	0.06	6.0659	0.04817
idade 2	1.60245				
Zona 1	2.33538	4.6	0.031	4.9237	0.02649
Nível socioeconómico 2	0.09178	10.5	0.015	11.475	0.009416
Nível socioeconómico 3	0.24586				
Nível socioeconómico 4	1.69962				
Funções 1	0.62964	8.3	0.04	9.3019	0.02554
Funções 2	-0.42865				
Funções 3	1.31266				
Transferido 1	-1.46496	2.5	0.11	2.6246	0.1052
Escala da dor 4	-1.75598	16.9	0.018	21.46	0.003146
Escala da dor 5	-2.55141				
Escala da dor 6	-2.24868				
Escala da dor 7	-3.02984				
Escala da dor 8	-3.5644				
Escala da dor 9	-3.9162				
Escala da dor 10	-4.23581				

Modelo final:

tempo total ~ idadecat + escala da dor +transferido+ zona + nível socioeconómico + funções.

Como pode ser observado na Tabela 7 os valores p, para os testes referentes à covariável transferido, levam a concluir que, apesar de escolhida pelo método *both stepwise,* ela

não é estatisticamente significativa. No entanto manteve-se a variável no modelo pois, de acordo com a literatura publicada, esta variável está associada ao modo como os doentes são encaminhados assim que chegam ao hospital de referência[7] e consequentemente influencia a variável resposta. Esta inclusão permitirá assim verificar de que modo a forma como os doentes são encaminhados no HSM, influencia o tempo total de isquémia para os doentes transferidos[42-43].

Método *backward stepwise*

O modelo obtido pelo método *backward stepwise* foi igual ao obtido pelo método *both stepwise*, pelo que não se apresentam aqui os resultados obtidos por este método.

Método *forward stepwise*

Os resultados obtidos pelo método de *forward stepwise* encontram-se na Tabela 8.

[7] Cada hospital pode ter uma forma diferente na recepção dos doentes que são transferidos, podendo passar ou não pelo SU.

Tabela 8: Modelo obtido a partir da aplicação do método *forward stepwise.*

Covariável	Estimativa coeficiente	Teste de Wald	Valor p (Wald)	*Deviance*	Valor p (Razão de Verosimilhança)
Idade 1	1.1058	4.2	0.12	4.4691	0.107039
Idade 2	1.3468				
Tipo de companhia 1	-1.9685	4.9	0.18	5.4104	0.1441
Tipo de companhia 2	-2.285				
Tipo de companhia 3	-0.3617				
Nível socioeconómico 2	0.1275	8.3	0.04	8.8185	0.0318
Nível socioeconómico 3	0.2384				
Nível socioeconómico 4	1.5101				
Escala da dor 4	-1.4889	15.7	0.028	18.967	0.00829
Escala da dor 5	-2.8543				
Escala da dor 6	-1.5292				
Escala da dor 7	-2.5874				
Escala da dor 8	-3.0555				
Escala da dor 9	-3.2769				
Escala da dor 10	-3.6038				
Funções 1	0.8704	5.9	0.12	6.5892	0.08621
Funções 2	-1.1735				
Funções 3	0.3338				

Modelo final:

tempo total ~ idadecat + tipo de companhia + nível socioeconómico + escala da dor + funções.

4.4.3 Comparação dos modelos obtidos

Após a aplicação dos diferentes métodos *stepwise* de selecção de covariáveis, *both, backward* e *forward* e ainda do método de selecção de "Hosmer e Lemeshow", obtiveram-se três modelos candidatos a modelos explicativos do tempo total de isquémia.Os modelos candidatos são resultantes dos métodos *both stepwise, forward stepwise* e "Hosmer e Lemeshow".

Os modelos obtidos com base no método *both stepwise* e no método *forward stepwise* estão aninhados no modelo obtido pelo método de "Hosmer e Lemeshow". O Teste de Razão de Verosimilhanças foi o método de selecção considerado para comparar os modelos referidos.

A primeira comparação foi feita entre o modelo obtido pelo método de "Hosmer e Lemeshow" (Modelo 1) e o modelo seleccionado pelo método *both stepwise* (Modelo 2). Os resultados são apresentados de seguida:

Modelo obtido pelo método de "Hosmer e Lemeshow"- Modelo 1:

Tempo total ~ idadecat + zona +transferido+ funções + nível socioeconómico + tipo de companhia + escala da dor.

Modelo obtido pelo método *both stepwise* - Modelo 2:

Tempo total ~ idadecat + zona + transferido +funções + nível socioeconómico + escala da dor.

```
Model 1: tempototal ~ idadecat + zona + funções + transferido +
nívelsócioeconómico + tipocompanhia + escalador
Model 2: tempototal ~ idadecat + escalador + zona + transferido +
nívelsócioeconómico + funções

  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       107     117.11
2       110     121.98 -3  -4.8719    0.1814
```

Com base no valor p, a diferença entre os dois modelos não é estatisticamente significativa pelo que o modelo obtido pelo método de *both stepwise* é o seleccionado.

De seguida foi comparado, o modelo obtido pelo método de "Hosmer e Lemeshow" com o modelo obtido pelo *forward stepwise*, que passará agora a ser designado por Modelo 2. Apresentam-se de seguida os resultados obtidos:

Modelo obtido pelo método *forward stepwise* - Modelo 2:

tempo total ~ idadecat + tipo de companhia + nível socioeconómico + escala da dor + funções.

```
Model 1: tempototal ~ idadecat + zona + funções + transferido +
nívelsócioeconómico + tipocompanhia + escalador
Model 2: tempototal ~ idadecat+ funções+nívelsócioeconómico
+ tipocompanhia + escalador

  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       109     121.53
2       107     117.11  2   4.4296    0.1092
```

Mais uma vez se verifica que a diferença entre os dois modelos não é estatisticamente significativa (valor p=0.1092), pelo que o modelo obtido pelo método de *forward stepwise* é o seleccionado.

Quanto à comparação entre os modelos obtidos pelo método *both stepwise* e pelo método *forward stepwise* usou-se o critério AIC uma vez que não estão aninhados. Os resultados são dados na Tabela 9.

Tabela 9: Valores obtidos para o critério AIC para cada um dos diferentes modelos escolhidos.

Método de selecção	Covariáveis no modelo	AIC
both stepwise	idadecat + escala da dor + zona+transferido+nível socioeconómico +funções	157.98
forward stepwise	Idadecat + tipo de companhia + escala da dor + nível socioeconómico + funções	159.53

Com base no critério AIC concluíu-se que o modelo obtido pelo método de ***both stepwise*** é o que melhor se ajusta ao conjunto de dados por possuir o menor valor para o AIC.

Modelo escolhido:

Tempo total ~ idadecat + escala da dor +transferido+ zona + nível socioeconómico + funções.

4.5 ANÁLISE DE RESÍDUOS

Depois de obtido o modelo final, que se segue:

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -0.20282 + 1.2883\ \text{idade1} + 1.60245\ \text{idade2} + 2.33538\ \text{zona1} + 0.09178\ \text{Nível socioeconómico 2}$$
$$+0.24586\ \text{Nível socioeconómico 3} + 1.69962\ \text{Nívelsocioeconómico 4} + 0.62964\ \text{Funções 1} - 0.42865\ \text{Funções 2}$$
$$+ 1.31266\ \text{Funções 3} - 1.46496\ \text{Transferido} - 1.75598\ \text{Escala da dor 4} - 2.55141\ \text{Escala da dor 5}$$
$$-2.24868\ \text{Escala da dor 6} - 3.02984\ \text{Escala da dor 7} - 3.5644\ \text{Escala da dor 8} - 3.9162\ \text{Escala da dor 9} - 4.23581\ \text{Escala da dor 10}$$

passou-se a avaliar a sua qualidade através da análise de resíduos, no que diz respeito à escolha da distribuição, da função de ligação e do preditor linear assim como observações mal ajustadas. Os resultados apresentam-se de seguida:

Adequação do preditor linear:

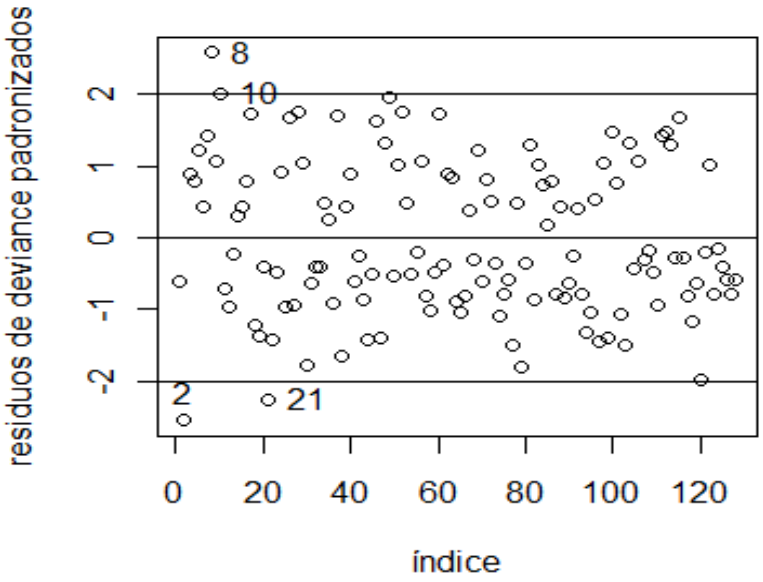

Figura 3: Gráfico dos resíduos desvio padronizados versus os índices das observações.

Na análise da Figura 5 observa-se um padrão nulo, em que os resíduos se encontram distribuídos em torno do zero com amplitude constante, e não há indícios de anomalias na escolha da função de ligação ou na escala de uma ou mais covariáveis. O gráfico permite-nos de igual forma identificar algumas observações consideradas *outliers*, nomeadamente a observação 2, 8, 10 e 21. Uma análise mais detalhada destes *outliers* revelou que os seus resíduos eram, -4.6845377, 4.9270805, 2.0048587 e -3.3019466, respectivamente. Estes valores encontram-se afastados do intervalo recomendado, -2 a 2. Para os indivíduos 2 e 21 estes valores devem-se ao facto de apresentarem características como, idade avançada, nível socioeconómico mais baixo e baixo nível na escala da dor, covariáveis que estariam teoricamente associadas um tempo total de isquémia aumentado, que não se verificou. Para o indivíduo 8 foi verificado o contrário, isto é, possuía características (individuo novo, nível socioeconómico mais alto, e um nível mais alto na escala da dor) que condicionavam uma diminuição do tempo de isquémia que não se verificou. Tal como para o individuo 8 o individuo 10 também não seria esperado um tempo aumentado uma vez que é também mais novo, e se encontrava a realizar algum tipo de esforço físico na altura dos sintomas.

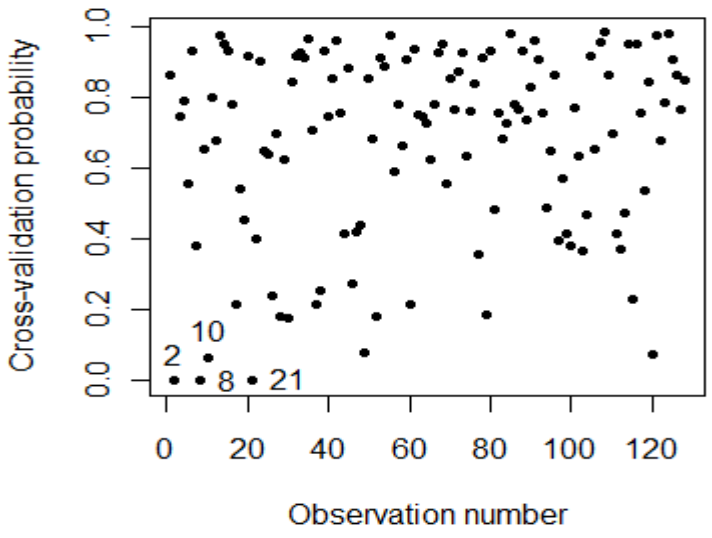

Figura 4: Gráfico das probabilidades cruzadas aproximadas para cada observação.

Na análise do FIgura 6, verifica-se novamente as mesmas observações, ou seja os indivíduos 2, 8, 10 e 21 são identificados como sendo *outliers*, apresentando baixa probabilidade de se realizarem, a saber probabilidades de 1.670339e-05, 5.223381e-06, 6.771117e-01 e 4.190028e-03 respectivamente.

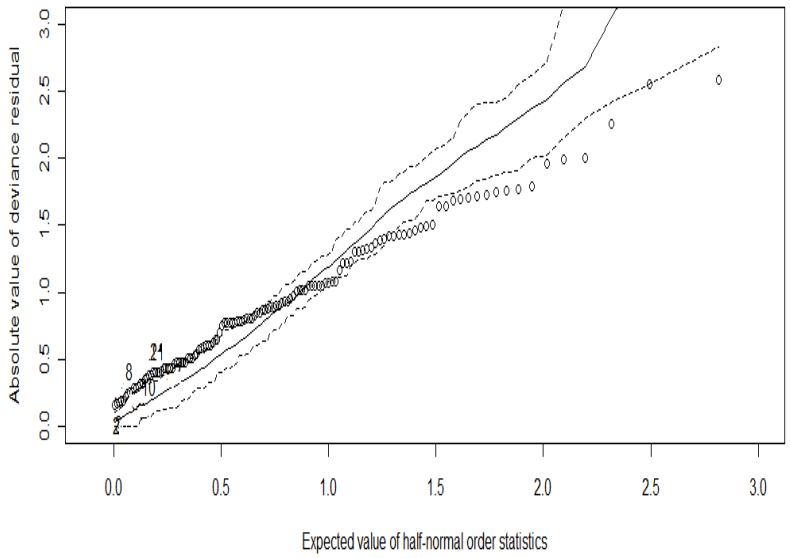

Figura 5: Gráfico *half-normal* dos resíduos desvio com o envelope usual.

Quanto à representação dos valores absolutos dos resíduos desvio num papel de probabilidades *half-normal* com o envelope usual (Figura 7), verifica-se que as observações 2, 8, 10 e 21, embora apresentem os maiores resíduos, encontram-se dentro do envelope o que indica que apesar de serem observações invulgares não possuem valores extremos o suficiente para serem consideradas como discordantes. No entanto observa-se um número elevado de valores fora do envelope.

Adequação da função de ligação:

Este método consiste em adicionar ao modelo, η^2, como covariável extra e examinar a mudança ocorrida no desvio através do Teste de Razão de Verosimilhanças:

Sendo que:

```
Model 1: tempototal ~ idadecat + escalador + zona+nívelsócioeconómico +
funções+transferido
Model 2: tempototal ~ idadecat2 + escalador + zona+nívelsócioeconómico +
funções +transferido+ eta2
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       110     121.98
2       109     120.81  1   1.1676    0.2799
```

Em que η^2 (`eta2`) consiste no preditor linear ao quadrado.

Como o valor p obtido para o Teste de Razão de Verosimilhanças não é significativo (valor p=0.2799) não existe evidência estatística para afirmar que a função de ligação não é adequada.

Identificação de observações influentes:

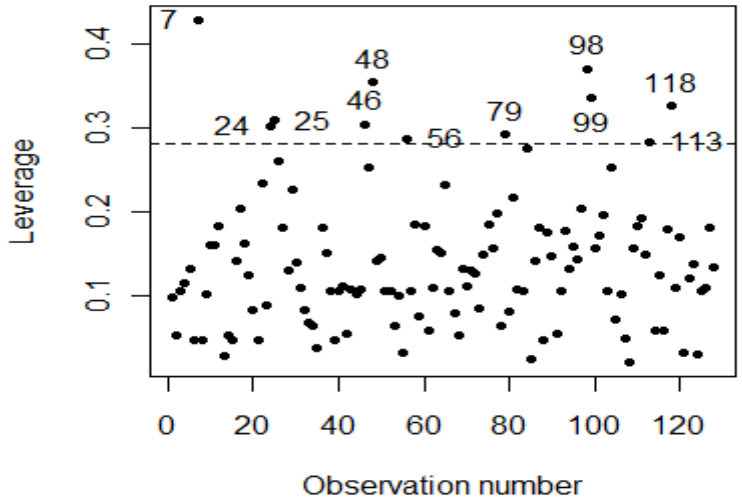

Figura 6: Gráfico do *leverage* em função do número de observação.

Na análise da Figura 8, são identificadas algumas observações como sendo influentes. Com excepção da observação 7, as restantes não apresentam valores suficientemente elevados para pôr em causa o modelo. Com o intuito de analisar com maior detalhe a influência desta observação, optou-se por realizar uma análise mais exaustiva, tendo-se feito o gráfico da estatística D (Figura 9) em função das observações, que de acordo com Collet[35] é o principal método para detectar observações influentes.

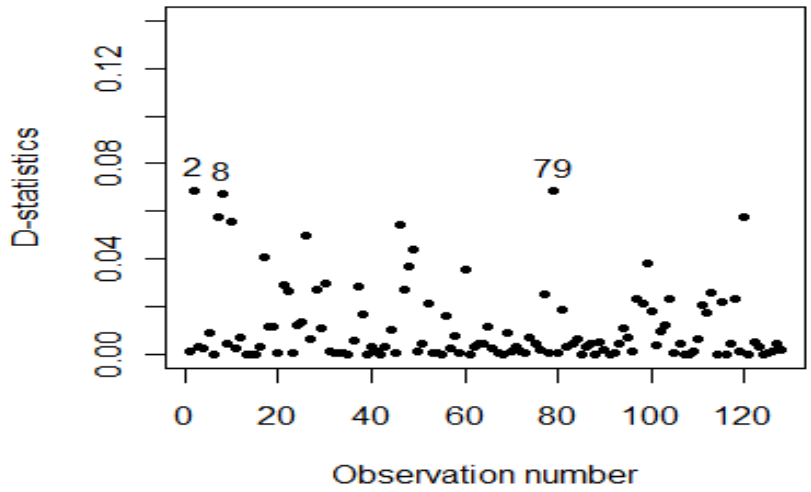

Figura 7: Gráfico da estatística D em função do número de observação.

Na Figura 9, as observações 2, 8 e 79 apresentam os maiores valores para a estatística D. Uma vez que as observações 2 e 8 foram também identificadas como sendo *outliers* e a observação 79 possui um valor muito elevado para a estatística D, optou-se por retirá-las do conjunto de dados e ajustar um novo modelo de forma a determinar se a sua exclusão produz alterações significativas nas estimativas ou leva mesmo à escolha de outras covariáveis.

Sobre este novo conjunto de dados foi aplicado o método de selecção *both stepwise*, que levou à obtenção de um novo modelo. Apresenta-se de seguida na Tabela 10, os valores dos coeficientes, da *Deviance* e do valor p do Teste de Razão de Verosimilhanças, para cada covariável incluída no novo modelo.

Tabela 10: Valores obtidos após aplicação do método *both stepwise.*

Covariável	Estimativa do coeficiente	*Deviance*	Valor p (Razão de Verosimilhança)
idadecat 1	1.67611	8.0704	0.01768
idadecat 2	2.18013		
Zona 1	3.25214	6.9418	0.008421
Nível socioeconómico 2	0.09343	12.442	0.006012
Nível socioeconómico 3	0.30526		
Nível socioeconómico 4	1.98776		
Funções 1	2.11979	23.953	0.0005331
Funções 2	-1.07389		
Funções 3	1.10109		
Transferido 1	-1.67155	2.8275	0.09266
Escala da dor 4	-1.77409	24.048	0.001118
Escala da dor 5	-2.61624		
Escala da dor 6	-1.74482		
Escala da dor 7	-3.51507		
Escala da dor 8	-3.77506		
Escala da dor 9	-4.06664		
Escala da dor 10	-4.82907		
Tipo de companhia 1	-2.87796	22.106	0.001159
Tipo de companhia 2	-2.68749		
Tipo de companhia 3	-0.27294		

O modelo referido na Tabela 10 contém todas as outras covariáveis previamente escolhidas e ainda foi acrescentada a covarariável tipo de companhia. A comparação entre o modelo

previamente seleccionado e o agora obtido, foi feita através do Teste de Razão de Verosimilhanças.

```
Model 1: tempototal ~ idadecat + escalador + transferido + zona +
nívelsócioeconómico + tipocompanhia + funções
Model 2: tempototal ~ idadecat2 + escalador + transferido + zona +
nívelsócioeconómico + funções
        Resid. Df Resid. Dev Df Deviance P(>|Chi|)
      1       104     99.251
      2       107    106.435 -3  -7.1833   0.06628 .
```

Para o nível de significância de 5% a diferença entre os dois modelos não é estatisticamente significativa pelo que se manteve o modelo inicial (Modelo 2).

Por fim foi utilizado o teste de Hosmer e Lemeshow para avaliar a:

Adequação do modelo:

Teste de Hosmer e Lemeshow:

```
$`Estatística Hosmer e Lemeshow`
[1] 6.66
$`P-valor`
[1] "0.574"
```

Para os níveis habituais de significância não se rejeita a hipótese de que o modelo se ajusta aos dados.

Com base na análise de resíduos considera-se que o modelo se adequa bem aos dados, mesmo tendo em conta os resultados apresentados no gráfico do papel de probabilidades *half-normal*.

4.6 PREDIÇÃO DO MODELO

Depois de escolhido o modelo e de se ter avaliado a qualidade do seu ajustamento ao conjunto de dados, passou-se a avaliar a capacidade de predição do mesmo, isto é, a capacidade de discriminar entre os indivíduos que de facto tiveram um tempo total aumentado dos que tiveram um tempo total normal. Para tal recorreu-se ao erro de predição e às curvas ROC, que se passa a apresentar.

Erro de predição:

Tabela 11: Tabela de contigência para valores observados e valores ajustados (cutpoint=0.5).

	Observados		
Ajustados	0	1	Total
0	60(0.469)	14(0.109)	74
1	14(0.109)	40(0.312)	54
Total	74	54	128

Ao cruzarem-se os valores observados para a variável resposta, com os valores ajustados superiores e inferiores a 0.5, obteve-se uma proporção de casos correctamente preditos de 78.12%. Entre parêntesis estão indicadas as percentagens sobre o total.

Curva ROC:

Para avaliar de forma mais precisa a capacidade de discriminação do modelo recorreu-se à curva ROC.

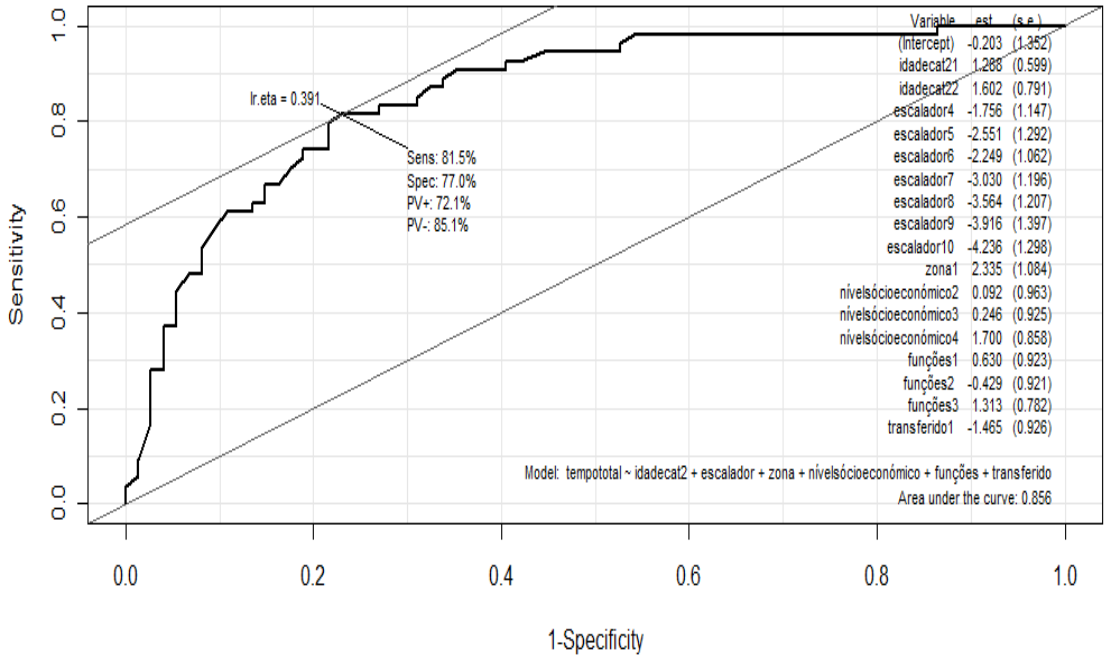

Figura 8: Curva ROC.

O valor obtido para a área sob a curva ROC (AUC) foi de 0.856, que de acordo com os valores propostos[34] indicam que o modelo possui uma capacidade de discriminação excelente. Verifica-se ainda que possui um valor de sensibilidade (81.5%) superior ao valor de especificidade (77.0%), ou seja, discrimina melhor os verdadeiros negativos (Y=0) do que os verdadeiros positivos (Y=1).

Capítulo 5

Resultados: Interpretação do modelo obtido

Depois de seleccionado o modelo e avaliado o seu ajuste ao conjunto de dados, procedeu-se à interpretação do mesmo. Na Tabela 12 estão indicadas as estimativas dos coeficientes do modelo e respectivos desvios padrão assim como o valor p referente ao teste de Wald, a estimativa do OR e a correspondente estimativa do IC ao nível de confiança 95%. Cada covariável incluída no modelo foi interpretada individualmente e por último foi feita a discussão do modelo e apresentadas as conclusões (Tabela 12).

Tabela 12: Valores obtidos para os OR e IC.

Covariável	Estimativa Coeficiente	Desvio padrão	Valor p	OR (ICi ,ICs)*
intercept	-0.203	1.352	0.881	0.816 (0.058,11.543)
idade1	1.288	0.599	0.032	3.627 (1.121,11.735)
idade2	1.602	0.792	0.043	4.965 (1.053-,23.420)
escala da dor 4	-1.756	1.147	0.126	0.173 (0.018,1.640)
escala da dor 5	-2.551	1.292	0.048	0.078 (0.006,0.981)
escala da dor 6	-2.249	1.062	0.034	0.106 (0.013,0.846)
escala da dor 7	-3.030	1.196	0.011	0.048 (0.005,0.503)
escala da dor 8	-3.564	1.207	0.003	0.028 (0.003,0.302)
escala da dor 9	-3.916	1.397	0.005	0.019 (0.001,0.308)
escala da dor 10	-4.236	1.298	0.001	0.014 (0.001,0.184)
zona1	2.335	1.085	0.031	10.333 (1.233,86.565)
nível socioeconómico 2	0.092	0.963	0.924	1.096 (0.166,7.236)
nível socioeconómico 3	0.246	0.925	0.790	1.278 (0.209,7.841)
nível socioeconómico 4	1.700	0.858	0.048	5.472 (1.019,29.390)
funções1	0.630	0.923	0.495	1.877 (0.307,11.467)
funções2	-0.429	0.921	0.642	0.651 (0.107,3.958)
funções3	1.313	0.782	0.093	3.716 (0.803,17.202)
transferido	-1.465	0.926	0.114	0.231 (0.038,1.419)

*IC para 95% de confiança.

5.1 INTERPRETAÇÃO

A interpretação de cada covariável foi feita com base no OR tendo em atenção que os indivíduos diferem apenas na característica de interesse partilhando os mesmos valores para as restantes covariáveis[34]. Foi igualmente dada, a título informativo, a estimativa a 95% de confiança do IC (Tabela 12), não foi feita a sua interpretação neste contexto dado a tratar-se de uma análise simples e que se tornaria repetitiva.

Idade:

A diferença entre a classe de referência (idade <55 anos) e as restantes classes (idade1 e idade2) é estatisticamente significativa, sendo o valor p de 0.032 e 0.043, respectivamente.

A ocorrência de ter um tempo total de isquémia aumentado (tempo total de isquémia superior a 6 horas) é 3.62 vezes superior em indivíduos com idades entre 55 e 75anos (idade1) e 4.96 vezes superior para indivíduos com idade superior aos 75 anos (idade2), quando comparados com os indivíduos com idade inferior a 55 anos (classe de referência). Foi também analisada a diferença entre as classes idade1 e idade2 e verificou-se que não é estatisticamente significativa (valor p= 0.642), sendo que a ocorrência de ter um tempo total de isquémia aumentado é 1.37 superior em indíviduos com idade superior aos 75 anos quando comparados com os indivíduos com idades entre 55 e 75 anos. (ver apêndice 4)

Escala da dor:

Tendo em atenção os valores p conclui-se que a diferença entre a classe de referência (indivíduos com um nível de dor 3) e a classe escala da dor 4 não é estatisticamente significativa, o que não acontece em relação à diferença entre a classe de referência e as restantes classes de escala da dor.

Verifica-se ainda que à medida que o nível de intensidade da dor aumenta na escala, existe uma redução da ocorrência de um tempo total de isquémia aumentado em relação ao nível de dor de referência. Tem-se assim, em relação à classe de referência, um decréscimo da ocorrência de um tempo total de isquémia aumentado de: 83% (não significativo) para o nível 4; 92% para o nível 5; 89% para o nível 6; 95% para o nível 7, 97% para o nível 8, 98% para o nível 9 e, por último, 98.5% para o nível 10.

Através da alteração da classe de referência para o nível 5 da escala da dor (por ser o nível médio de dor para esta escala), verificou-se que a diferença entre esta e as restantes classes não é estatisticamente significativa. Com excepção do nível de dor 4 e nível de dor 6, para os quais foi observado um aumento na ocorrência de um tempo total de isquémia aumentado de 2.22 e 1.35, respectivamente, os restantes níveis de dor apresentaram uma redução na ocorrência de um tempo total de isquémia aumentado, de 38% para o nível 7, 64% para o nível 8, 74% para o nível 9 e 81% para o nível 10.

O resultado das comparações entre as diferentes classes da escala da dor estão representadas no apêndice 4.

Zona:

A ocorrência de ter um tempo total de isquémia aumentado é 10 vezes superior em indivíduos com proveniência de uma zona rural em relação aos provenientes de uma zona urbana (classe de referência), sendo a diferença entre as duas zonas estatisticamente significativa (valor p=0.031).

Nível socioeconómico:

Em relação a esta variável, a classe de referência definida é o nível socioeconómico alto e verifica-se que a diferença entre esta e as classes correspondentes aos níveis socioeconómico médio alto (2) e médio baixo (3) não é estatisticamente significativa (valores p 0.924 e 0.790, respectivamente). Em relação ao nível socioeconómico baixo (4) essa diferença é estatisticamente significativa (valor p = 0.048). À medida que o nível socioeconómico diminui, verifica-se um aumento da ocorrência de um tempo total de isquémia aumentado em relação à classe de referência. Esse aumento é 1.09 vezes superior para os sujeitos de nível socioeconómico médio alto (2), 1.27 superior para os indivíduos de nível socioeconómico médio baixo (3) e 5.47 vezes superior para os de nível socioeconómico baixo (4).

Quando alterada a classe de referência para o nível socioeconómico baixo (4), verifica-se que a diferença entre esta e as classes, nível socioeconómico médio baixo (3) e a classe nível socioeconómico médio alto (2), são estatisticamente significativas, valor p = 0.01639

e valor p= 0.01059, respectivamente. Com o aumento do nível socioeconómico observa-se uma redução na ocorrência de um tempo total de isquémia aumentado. Observa-se uma diminuição de 76.63% para os indivíduos de nível socioeconómico médio baixo (3) e uma diminuição de 79.97% para indíviduos nível socioeconómico médio alto (2), em relação à classe de referência.

A diferença entre a classe nível socioeconómico médio baixo (3) e a classe nível socioeconómico médio alto (2), não é estatisticamente significativa (valor p= 0.82170). Verificando-se uma redução de 14.28%, na ocorrência de um tempo total de isquémia aumentado para os indíviduos do nível socioeconómico médio alto (2). (ver apêndice 4)

Funções:

A diferença entre a classe de referência, indivíduos que praticavam uma actividade de lazer, e as classes designadas por funções1 (indivíduos que se encontravam a trabalhar) e funções2 (indivíduos que se encontravam a realizar algum esforço físico) não é estatisticamente significativa (valores p 0.495 e 0.642, respectivamente). Pelo contrário, em relação à classe designada por função3 (indivíduos que se encontravam a dormir) verifica-se uma diferença estatisticamente significativa (valor p=0.093), com a ocorrência de um tempo total de isquémia aumentado de 3.71 vezes superior nesses indivíduos em relação à classe de referência. Em relação aos indivíduos que praticavam uma actividade de lazer (classe de referência) a ocorrência de um tempo total de isquémia aumentado é 1.87 superior nos indivíduos que se encontravam a trabalhar. Nos indivíduos que se encontravam a realizar algum esforço físico, verifica-se uma redução de 34.8% na ocorrência de um tempo total de isquémia.

Quando alterada a classe de referência para funções1, observou-se que as diferenças entre esta e as classes funções2 e funções3 não eram estatisticamente significativas, valor p= 0.21619 e valor p=0.3449 respectivamente. Apresentando uma ocorrência de um tempo total de isquémia aumentado 2.88 superior para a classe classe funções2 e 1.98 superior para a classe funções3.

No entanto a diferença entre as classes funções2 e a classe funções3 é estatisticamente significativa, valor p= 0.0086. A ocorrência de um tempo total de isquémia aumentado é

5.70 vezes superior para a classe funções3 quando comparada com a classe de referência. (ver apêndice 4)

Transferido:

Por último, observa-se uma diminuição de 76.8% (não significativa, valorp=0.114) da ocorrência de um tempo total de isquémia aumentado nos sujeitos que foram transferidos de outro hospital.

5.2 DISCUSSÃO E CONCLUSÃO

O SCA mais especificamente o seu subtipo clínico designado por EAMEST, é a manifestação mais grave do SCA cujo prognóstico é muito variável, podendo ser mais ou menos favorável, dependendo de vários factores, que são essencialmente o seu diagnóstico e tratamento precoces.

Durante as últimas décadas, aprendeu-se que, no que diz respeito ao SCA, o tempo significa miocárdio viável. O significado por trás desta afirmação é que, quanto mais cedo o tratamento é iniciado, maior a probabilidade de que os danos do miocárdio sejam limitados e que a função miocárdica seja mantida[44]. Da mesma forma os resultados do tratamento do SCA, dependem do tempo que decorreu entre o início dos sintomas e a revascularização da artéria responsável pelo EAM[45].

Uma vasta quantidade de estudos apontam para que, os factores associados ao aumento do **tempo total de isquémia miocárdica** na sua maioria, são inerentes ao doente[39, 44, 46].

A utilização da regressão logística na análise dos dados desta coorte, permitiu a obtenção de um modelo, onde foram identificados factores implicados na ocorrência de um tempo total de isquémia.

No caso do modelo obtido foram identificadas seis covariáveis (idade do doente, nível de intensidade da dor, zona de proveniência, nível socioeconómico, funções que se encontrava a realizar, transferido) associadas à variável resposta (tempo total de isquémia).

Cada uma dessas covariáveis apresenta várias classes/categorias, associadas à variável resposta, que se passa a enunciar:

(i) estão associadas a um aumento da ocorrência do tempo total de isquémia superior a 6 horas as categorias: idade avançada, baixa intensidade da dor, zona de proveniência rural, baixo nível socioeconómico, estar a trabalhar ou estar a dormir;

(ii) estão associadas a uma diminuição da ocorrência do tempo de isquémia superior a 6 horas as categorias: ter sido transferido de outro hospital, encontrar-se a fazer um "esforço físico".

As covariáveis apresentadas no modelo logístico obtido estão de acordo com os resultados indicados nos diversos estudos internacionais realizados nesta área[44, 46]. A maioria das covariáveis, à excepção da idade, são passíveis de serem modificadas, contribuindo para a redução do tempo total de isquémica.

De forma mais detalhada e com base na revisão da literatura, apresenta-se para cada covariável as possíveis explicações que justificam o resultado obtido neste modelo logístico, ou seja a sua associação ao tempo de isquémia aumentado.

Idade: Aumento da idade

Verifica-se de forma consistente na literatura a existência de um aumento do tempo total de isquémia relacionado com o aumento da idade[47]. Esta associação é resultado da imprecisão dos sintomas, devido a uma incerteza causada pela apresentação de sintomas menos comuns como dor no peito menos acentuada e presença de sintomas de outras patologias presentes no doente, que podem mascarar os sintomas inerentes ao SCA.

Uma das explicações propostas baseia-se no facto das pessoas de idade avançada passarem a aceitar a presença de sintomas novos como parte da vida, associada ao facto de que identificar a origem exacta da sintomatologia e agir rapidamente no sentido de procurar ajuda médica em tempo útil pode se tornar mais difícil com o aumento da idade[48].

Escala da dor: dor de menor intensidade

Neste estudo observou-se que, valores baixos na escala da intensidade da dor condicionam um maior atraso no ínicio do tratamento, quando comparados com os factores sociodemográficos. Isto poderá dever-se à incongruência entre as expectativas dos doentes relacionadas com os sintomas e a experiência real, como já foi verificado noutros estudos[48].

Muitos doentes relataram o facto de se sentirem desiludidos com a intensidade dos sintomas, nomeadamente da dor, pois não se assemelham a um típico "enfarte de Hollywood", como visto na televisão. Uma evolução lenta e cumulativa dos sintomas, com menor intensidade da dor, produz ainda maior demora[48].

Nível socioeconómico: nível socioeconómico baixo

Nos países em desenvolvimento, o SCA tem sido historicamente descrito como mais comum nos indivíduos com nível socioeconómico mais alto, facto que se inverteu nas últimas décadas[49]. Estudos realizados em países desenvolvidos sugerem que o baixo nível socioeconómico está associado a uma maior incidência de SCA e mortalidade associada. Estes factos devem-se à maior prevalência de factores de risco para doenças cardíacas (pressão arterial alta, tabagismo e diabetes) e ao menor uso de medicamentos, bem como a redução na adesão e no acesso rápido ao tratamento que se verificam nesta classe socioeconómica[49].

A OMS refere que as vias pelas quais a condição socioeconómica pode afectar as doenças cardiovasculares incluem: o estilo de vida e padrões de comportamento, facilidade de acesso aos cuidados de saúde e stress crónico[2].

Zona: zona de proveniência rural/distância do centro de tratamento

A necessidade de percorrer longas distâncias para chegar ao hospital, nomeadamente o facto de residir numa zona rural, estão associados a maiores atrasos[48].

Funções que desempenhava: repouso ou diferentes actividades físicas

Estudos demonstram que aqueles indivíduos que se encontravam a descansar ou a dormir no início dos sintomas demoram mais tempo do que aqueles que se encontravam a exercer algum tipo de actividade física.

Os compromissos sociais podem prevalecer sobre o impulso de procurar cuidado imediato, até mesmo para os sintomas agudos. As situações e as circunstâncias podem restringir o comportamento do indivíduo[46, 48].

Transferido: Transferido de outro hospital para o centro de tratamento

Ao contrário da maioria dos estudos, que verificam existir um aumento do tempo total de isquémia nos doentes transferidos, neste estudo verificamos uma diminuição da ocorrência de aumento do tempo nestes doentes. Isto poderá ser devido ao facto de os doentes que são transferidos não passarem pelo SU e seguirem directamente para o laboratório de hemodinâmica onde é realizado a ICP primária, o que poderá estar associado a uma diminuição do tempo total de isquémia[42].

Um outro aspecto importante do modelo de regressão logística é a capacidade deste em discriminar os sujeitos com tempo total de isquémia aumentado, daqueles com tempo total de isquémia dentro do intervalo considerado recomendado. De acordo com a classificação teórica, o modelo obtido apresenta uma excelente capacidade discriminatória, mas que só a prática clínica poderá confirmar. Neste momento, não se dispõe de nenhum conjunto de novos dados que permita validar externamente essa capacidade.

Em conclusão, a aplicação conjunta dos conhecimentos de Biologia e de Estatística (Bioestatística) responderam à questão científica levantada pela prática clínica diária dos técnicos de saúde: Quais os factores associados ao aumento do tempo total de isquémica no SCA? Este trabalho permitiu identificar covariáveis associadas ao tempo total de isquémia e possibilitou a selecção de aquelas que são passíveis de serem modificadas para optimização da terapêutica nestes doentes, dado que identificou os doentes que constituem um grupo de alto risco, para os quais devem ser dirigidos os esforços educacionais. Em particular, os doentes que devem receber instruções sobre os sinais de

alerta, nomeadamente a severidade dos sintomas que devem levá-los a procurar precocemente os cuidados de saúde após o início dos sintomas e a forma de activação rápida da rede emergência médica específica para o SCA.

Bibliografia

1. Allender S, S.P., Peto V, Rayner M, Leal J, Luengo-Fernandez R,, *European cardiovascular disease statistic*. 2008, Department of Public Health, University of Oxford,: Oxford.

2. World health organization. *Mortality country fact sheet 2006* 2009 2008 November 20 [cited 2011.

3. Braunwald K, H.F., Jameson L. Harrison *Princípios de medicina interna*. 16 ed. Vol. 1. 2006, Madrid: Mcgerawhill interamericana.

4. Thygesen K, A.J., White HD., *Universal Definition of Myocardial Infarction. Circulation.* Journal of the American Heart Association, 2007.

5. Instituto Nacional de Estatística (INE). *Inquérito Naciona de Saúde*. 2009 [cited 2012; Available from: http://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=69365215&PUBLICACOESmodo=2.

6. Sociedade Portuguesa de Cardiologia. *Registo Nacional de Síndromes Coronários Agudos. Número de Registos Recebidos no CNCDC,*. [cited 2011; Available from: http://www.spc.pt/cncdc/.

7. Jacobs, A.K., *Regional systems of care for patients with ST-elevation myocardial infarction: being at the right place at the right time.* Circulation, 2007. **116**(7): p. 689-92.

8. Jeronimo Sousa, P., et al., *Primary PCI in ST-elevation myocardial infarction: Mode of referral and time to PCI.* Rev Port Cardiol, 2012.

9. Costa S, C.D., Romeira H, Gomes Lopes J, Baeta C,, *Intervenção Coronária Percutânea Primária em doentes com Síndrome Coronário Agudo com elevação do segmento ST: Referenciação versus não referenciação,*. Cardiopulmonar Associação Portuguesa de Cardiopneumologia,, 2008.

10. Seeley, S., Tate *Anatomy and Physiology*, ed. t. edition. 2005: McGraw-Hill Higher Education.

11. Alexander RW, S.R., Fuster V, O'Rourke RA, Roberts R, Sonnenblick EH. Hurst´s *O coração*. 10th edition ed. 2002, Lisboa: McGraw Hill.

12. Braunwald E, Z.D., Libby P,, *Heart Disease – a textbook of cardiovascular medicine*. 6th edition ed. 2001: Philadelphia Saunders,.

13. Kong D, B.M., Connor C,, *Progressos na Abordagem das Síndromes Coronárias Agudas.* Vol. volume 6. 2002 Hospital Practice.

14. Libby, P., *Current concepts of the pathogenesis of the acute coronary syndromes.* Circulation, 2001. **104**(3): p. 365-72.

15. Théroux P, *Acute coronary syndromes – a companion to Branwald's heart disease.* . 2003: Philadelphia Saunders.

16. Brodie, B.R., et al., *Door-to-balloon time with primary percutaneous coronary intervention for acute myocardial infarction impacts late cardiac mortality in high-risk patients and patients presenting early after the onset of symptoms.* J Am Coll Cardiol, 2006. **47**(2): p. 289-95.

17. Brokalaki, H., et al., *Factors associated with delayed hospital arrival among patients with acute myocardial infarction: a cross-sectional study in Greece.* Int Nurs Rev, 2011. **58**(4): p. 470-6.

18. Williams, W.L., *Guidelines to reducing delays in administration of thrombolytic therapy in acute myocardial infarction.* Drugs, 1998. **55**(5): p. 689-98.

19. Sousa P, L.A., Santiago H,, *Paradigma dos Tempos na Angioplastia Primária.* Revista Portuguesa de Cardiologia, 2001.

20. Westerhout, C.M., et al., *The influence of time from symptom onset and reperfusion strategy on 1-year survival in ST-elevation myocardial infarction: a pooled analysis of an early fibrinolytic strategy versus primary percutaneous coronary intervention from CAPTIM and WEST.* Am Heart J, 2011. **161**(2): p. 283-90.

21. Antman, E.M., et al., *ACC/AHA guidelines for the management of patients with ST-elevation myocardial infarction; A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Revise the 1999 Guidelines for the Management of patients with acute myocardial infarction).* J Am Coll Cardiol, 2004. **44**(3): p. E1-E211.

22. McNamara, R.L., et al., *Effect of door-to-balloon time on mortality in patients with ST-segment elevation myocardial infarction.* J Am Coll Cardiol, 2006. **47**(11): p. 2180-6.

23. Ferreira D, *Estratégias de Reperfusao no Tratamento do Enfarte Agudo do Miocárdio. Transporte para centros de referência .* Revista Portuguesa de Cardiologia 2007.

24. Rogers, W.J., et al., *Treatment and outcome of myocardial infarction in hospitals with and without invasive capability. Investigators in the National Registry of Myocardial Infarction.* J Am Coll Cardiol, 2000. **35**(2): p. 371-9.

25. Herrmann, H.C., *Transfer for primary angioplasty: the importance of time.* Circulation, 2005. **111**(6): p. 718-20.

26. Dobson, A.J., *An introduction to generalized linear models*. 2nd ed ed. Chapman & Hall/CRC texts in statistical science series. 2002, Boca Raton, Fla., London: Chapman & Hall/CRC. vii, 225 p.

27. Venables, W.N., B.D. Ripley, and MyiLibrary., *Modern applied statistics with S*, in *Statistics and computing*. 2002, Springer: New York. p. xi, 495 p.

28. M. Antónia Amaral Turkman, G.L.S., *Modelos Lineares Generalizados-da teoria a prática-*. 2000, Lisboa: Edições SPE.

29. Lehman, E.L., *Testins statistical hypotheses* ed. n. edition. 1986, New York: Wiley.

30. Azzalini A, *Statistical Inference: based on the likelihood.* 1996, London: Champmam and Hall,.

31. Fahrmeir, L., Kaufmann, H, , *Consitency and Asymptotic Normality of the Manximum Likelihood Estimator in Generalized Linear Model.Annals of Statistics.* 1985.

32. McCullagh, P. and J.A. Nelder, *Generalized linear models*. 2nd ed. Monographs on statistics and applied probability 37. 1989, London: Chapman and Hall. xix, 511 p.

33. McCullagh, P. and J.A. Nelder, *Generalized linear models*. Monographs on statistics and applied probability. 1983, London: Chapman and Hall. xiii,261p.

34. Hosmer, D.W. and S. Lemeshow, *Applied logistic regression*. 2nd ed ed. Wilet series in probability and statistics, texts and references section. 2000, New York, Chichester: Wiley. xii, 373 p.

35. Collett, D., *Modelling binary data*. 2nd ed ed. 2003, Boca Raton, London: Chapman & Hall. 387 p.

36. Felgueiras, M.M., *Análise de dados binários*, in *Faculdade de Ciências*. 2003, Universidade de Lisboa: Lisboa.

37. Direcção Geral da Saúde, *A Dor como 5º sinal vital. Registo sistemático da*

intensidade da Dor, Ministério da Saúde, Editor. 2003: Lisboa.

38. Ruston, A., J. Clayton, and M. Calnan, *Patients' action during their cardiac event: qualitative study exploring differences and modifiable factors.* BMJ, 1998. **316**(7137): p. 1060-4.

39. Ting, H.H., et al., *Factors associated with longer time from symptom onset to hospital presentation for patients with ST-elevation myocardial infarction.* Arch Intern Med, 2008. **168**(9): p. 959-68.

40. Bernardi, G., *[Relationship of symptom-onset-to-balloon time and door-to-balloon time with mortality in patients undergoing angioplasty for acute myocardial infarction].* Ital Heart J Suppl, 2000. **1**(11): p. 1485-7.

41. Agresti, A., *Categorical data analysis*. Wiley series in probability and mathematical statistics. 2002, New York: Wiley. xv, 558 p.

42. Miedema, M.D., et al., *Causes of delay and associated mortality in patients transferred with ST-segment-elevation myocardial infarction.* Circulation, 2011. **124**(15): p. 1636-44.

43. Pinto, D.S., et al., *Benefit of transferring ST-segment-elevation myocardial infarction patients for percutaneous coronary intervention compared with administration of onsite fibrinolytic declines as delays increase.* Circulation, 2011. **124**(23): p. 2512-21.

44. Herlitz, J., et al., *Factors of importance for patients' decision time in acute coronary syndrome.* Int J Cardiol, 2010. **141**(3): p. 236-42.

45. Walkiewicz, M., et al., *Acute coronary syndrome--how to reduce the time from the onset of chest pain to treatment?* Kardiol Pol, 2008. **66**(11): p. 1163-70; discussion 1171-2.

46. Sullivan, M.D., et al., *Understanding why patients delay seeking care for acute coronary syndromes.* Circ Cardiovasc Qual Outcomes, 2009. **2**(3): p. 148-54.

47. Moser, D.K., et al., *Gender differences in reasons patients delay in seeking treatment for acute myocardial infarction symptoms.* Patient Educ Couns, 2005. **56**(1): p. 45-54.

48. Moser, D.K., et al., *Reducing delay in seeking treatment by patients with acute coronary syndrome and stroke: a scientific statement from the American Heart Association Council on cardiovascular nursing and stroke council.* Circulation, 2006. **114**(2): p. 168-82.

49. Moser, D.K., et al., *Reducing delay in seeking treatment by patients with acute coronary syndrome and stroke: a scientific statement from the American Heart Association Council on Cardiovascular Nursing and Stroke Council.* J Cardiovasc Nurs, 2007. **22**(4): p. 326-43.

Apêndice 1

Questinário 1

Questionário 1

Dados retirados dos processos clínicos:

1- Dados pessoais:
Nome:_____

Número Hospitalar:_____

2- Raça: Caucassiana __ Negra__Asiática__

3- Sexo: F__M__

4- Dados do doente na altura da ocorrência de interesse:

5- Data de ocorrência do SCA__/__/2010

6- Transferido de outro hospital: Sim__ Não__

7- Zona onde se encontrava: Rural__ Urbana__

8- Idade:__

9- Antecedentes de SCA: Sim__ Não__

10- Factores de Risco: Sim__Não__

11- Hora de início da sintomatologia: __:__

12- Hora de activação dos sistemas de emergência (caso aplicável):__:__

13- Hora de entrada no SU: __:__

14- Hora de restabelecimento do fluxo coronária: __:__

15- Estado do doente, após tratamento: Óbito: __ Alta: __

Apêndice 2

Questionário 2

Questionário 2

Dados obtidos através do contacto telefónico

- Data do contacto telefónico: __ /__/___

1- Nível socioeconómico/Escala de Graffar (ver anexo I):

Classe I__ Classe II__ Classe III__ Classe IV__ Classe V__

2- Na altura do início dos sintomas, encontrava-se acompanhado?
Sim__Não__

2.a) Se sim por:

Familiares__ amigos__ colegas de trabalho__

3- Que funções se encontrava a desempenhar na altura do início dos sintomas?

4- Tinha conhecimento sobre o que é o enfarte do miocárdio?
Sim__ Não__

5- Numa escala de 1 até 10, correspondendo 1 a ausência de dor e 10 a dor máxima, como classificaria a dor que experimentou na altura?

6- Óbito pós-Alta: Sim__ Não__
6.a) No caso de sim, causas:

7- Recidiva de EAM: Sim__Não__

8.a) No caso de sim, tipo de tratamento:
Angioplastia:__ Cirurgia__

8- Incidência de AVC: Sim__Não__

9- Novamente hospitalizado: Sim__Não__

10.a) No caso de sim, Causa_____

Apêndice 3

Introdução da variável idade como contínua.

Introducção da interacção: idade contínua versus escala da dor.

Tendo em atenção a importância clínica que a interacção entre a idade e a escala da dor representa, optou-se por utilizar a covariável idade como contínua passando a aplicar diferentes métodos de selecção de covariáveis.

São de seguida apresentados os resultados obtidos para a análise referida.

Método de selecção: "Hosmer e Lemeshow".

Passo 1: Construção de modelos univariados

Tabela 1: Modelos simples.

Covariável	Coeficiente	Teste de Wald	Valor p pelo teste de Wald	*Deviance*	Valor p pelo teste de razão de verossimilhança
Idade	0.03246	5.3678	0.02051	5.668	0.01728
Sexo	-0.01307	0.001	0.9745	0.0010214	0.9745
Dia 1	-0.3424	0.7734	0.3792	0.78255	0.3764
Transferido 1	0.7469	2.1936	0.1386	2.2277	0.1356
Zona 1	1.2528	3.947	0.04695	4.3141	0.0378
Antecedentes 1	0.1606	0.1161	0.7333	0.11569	0.7338
Nível socioeconómico 2	0.1054	11.673	0.00859	12.513	0.005818
Nível socioeconómico 3	0.6444				
Nível socioeconómico 4	1.5787				
Nível escolaridade 2	-0.8303	9.0806	0.01067	9.9828	0.006796
Nível escolaridade 3	-1.4365				
Tipo de companhia 1	-1.8192	11.5	0.009308	14.426	0.002379
Tipo de companhia 2	-2.2892				
Tipo de companhia 3	-0.6893				
Funções 1	0.6337	7.5319	0.05675	8.0941	0.04411
Funções 2	0.4055				
Funções 3	1.3672				
Conhecimento da doença 1	-0.9484	6.6445	0.009946	6.8303	0.008962
Escala da dor 4	-0.2513	18.166	0.01124	21.313	0.003334
Escala da dor 5	-1.6094				
Escala da dor 6	-0.9555				
Escala da dor 7	-2.0149				
Escala da dor 8	-1.8608				
Escala da dor 9	-2.8904				
Escala da dor 10	-2.3979				

Foram seleccionadas as variáveis com valor p inferior a 0.25. Após a aplicação deste critério foram seleccionadas as seguintes covariáveis:

- Idade
- Transferido
- Zona
- Nível socioeconómico
- Nível escolaridade
- Tipo de companhia
- Funções
- Conhecimento da doença
- Escala da dor.

Uma vez identificadas as possíveis variáveis explicativas passamos a construção do modelo (Model 1) contendo todas as variáveis seleccionadas. Mais uma vez são apresentados os valores para os coeficientes, o valor para os testes Wald e os valores p resultantes destes (Tabela 2).

Passo 2: Construção do modelo múltiplo contendo as covariáveis selecionadas a partir dos modelos simples ajustados.

Tabela 2: Modelo múltiplo obtido após análise dos modelos simples.

Covariável	Coeficientes	Teste de Wald	Valor p de Wald	*Deviance*	Valor p para o teste de Razão de Verosimilhanças
Idade	0.03043	2.4	0.12	2.451	0.1174
Transferido 1	-1.30453	2.1	0.15	2.1827	0.1396
Zona 1	2.14391	3.7	0.054	3.9323	0.04737
Nível socioeconómico 2	0.22862	4.9	0.18	5.2789	0.1525
Nível socioeconómico 3	0.04264				
Nível socioeconómico 4	1.50640				
Nível escolaridade 2	0.52107	0.61	0.74	0.60905	0.7375
Nível escolaridade 3	-0.08025				
Tipo de companhia 1	-1.97705	4.8	0.19	5.2551	0.154
Tipo de companhia 2	-2.10823				
Tipo de companhia 3	-0.29807				
Funções 1	1.14413	6.4	0.093	7.0483	0.07037
Funções 2	-0.87574				
Funções 3	0.60570				
Conhecimento da doença 1	-0.41443	0.57	0.45	0.57609	0.4478
Escalador da 4	-0.85976	15.5	0.03	19.033	0.008085
Escalador da 5	-2.18723				
Escalador da 6	-1.30174				
Escala da dor 7	-1.91293				
Escala da dor 8	-2.61600				
Escala da dor 9	-3.34171				
Escala da dor10	-3.51992				

De seguida foram analisados os valores obtidos para cada covariável tendo em especial atenção o valor para o teste de Wald e o respectivo valor p, assim como também devem ser comparados os coeficientes obtidos para cada covariável com os coeficientes obtidos pelos modelos simples. Sendo assim foram seleccionadas as seguintes variáveis:

- Idade,
- Transferido,
- Zona,
- Funções,
- Nível socioeconómico,
- Tipo de companhia,
- Escala da dor,

Tendo em conta que as covariáveis referidas anteriormente, apresentavam valor $p<0.25$ e não se observaram grandes discrepâncias nas estimativas dos coeficientes obtidos, um novo modelo múltiplo foi ajustado.

Passo 3: Criação de um novo modelo de regressão múltipla com as covariáveis selecionadas do modelo ajustado anteriormente

Tabela 3: Modelo múltiplo após selecção de covariáveis.

Covariável	Coeficiente	Teste de Wald	Valor p de Wald	*Deviance*	Valor p para o teste de Razão de verosimilhanças
Idade	0.15210	2.1	0.15	2.0864	0.1486
Zona 1	0.06661	3.4	0.067	3.5757	0.05863
Funções 1	0.35044				
Funções 2	0.33414	6.5	0.088	7.1961	0.0659
Funções 3	0.63806				
Transferido 1	0.16787	1.9	0.17	1.9677	0.1607
Nível socioeconómico2	0.65293				
Nível socioeconómico3	0.58436	9.0	0.029	9.7302	0.021
Nível socioeconómico4	0.03354				
Tipo companhia1	0.04349				
Tipo companhia2	0.12818	5.0	0.17	5.5373	0.1364
Tipo companhia3	0.66365				
Escala da dor 4	0.30506				
Escalada dor 5	0.05342				
Escala da dor 6	0.16941				
Escala da dor 7	0.04889	16.2	0.024	19.968	0.00564
Escala da dor 8	0.01323				
Escala da dor 9	0.00952				
Escala da dor10	0.00294				

Depois de obtido dois modelos múltiplos estes devem ser comparados, de forma a escolher o que se ajusta melhor aos dados. Os modelos foram comparados recorrendo ao Teste de Razão de Verosimilhanças apresentado logo de seguida:

```
Model 1: tempototal ~ idade + transferido + zona + nívelsócioeconómico +
nivelescolaridade2 + tipocompanhia + funções + conhecimentodoença +
escalador
Model 2: tempototal ~ idade + zona + funções + transferido +
nívelsócioeconómico + tipocompanhia + escalador
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       105     119.33
2       108     120.39 -3  -1.0661    0.7853
```

Com base no valor p (0.7853) obtido pelo Teste de Razão de Verosimilhanças a diferença entre os dois modelos não é estatisticamente significativa (aos níveis usuais de significância) pelo que se deve escolher o modelo com menos covariáveis, ou seja o:

Model 2: Tempo total ~ idade + transferido + zona + funções + nível socioeconómico + tipo de companhia + escala da dor

Uma vez que neste último modelo (Model 2) todas as covariáveis são significativas e não há grandes discrepâncias nas estimativas dos coeficientes obtidos, de acordo com Hosmer e Lemeshow mantém-se assim todas as covariáveis (Model 2). Ainda de acordo com Hosmer e Lemeshow, foi introduzido no modelo (Model 2) cada uma das variáveis que tinham sido dispensadas na análise univariada, uma vez que podem passar a ser significativas quando controladas por outras covariáveis. Os resultados não são aqui apresentados, uma vez que nenhuma das covariáveis acrescentadas no modelo se verificou significativa, nem causou discrepâncias nos coeficientes.

Linearidade:

É necessário ainda, nesta fase, verificar a linearidade da covariável idade na escala logit, para isto foi ajustado um modelo linear generalizado aditivo que contempla um termo não linear (spline cúbico), associado à covariável idade. Por último, este modelo deve ser comparado, pelo Teste de Razão de Verosimilhanças, ao modelo que contém apenas termos lineares.

```
Model 1: tempototal ~ idade + zona + funções + transferido +
nívelsócioeconómico + tipocompanhia + escalador
Model 2: tempototal ~ idade + s(idade) + escalador + transferido +
tipocompanhia + nívelsócioeconómico + funções + zona
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       108     120.39
2       105     118.65  3   1.7475    0.6264
```

Com base no valor p (0.6264) obtido a diferença entre os dois modelos não é estatisticamente significativa pelo que se pode concluir que não se rejeita a hipótese da linearidade da idade na escala logit.

Na Figura 1 está representado o gráfico da *idade* vs s(*idade*) onde se confirma a hipótese da existência de linearidade.

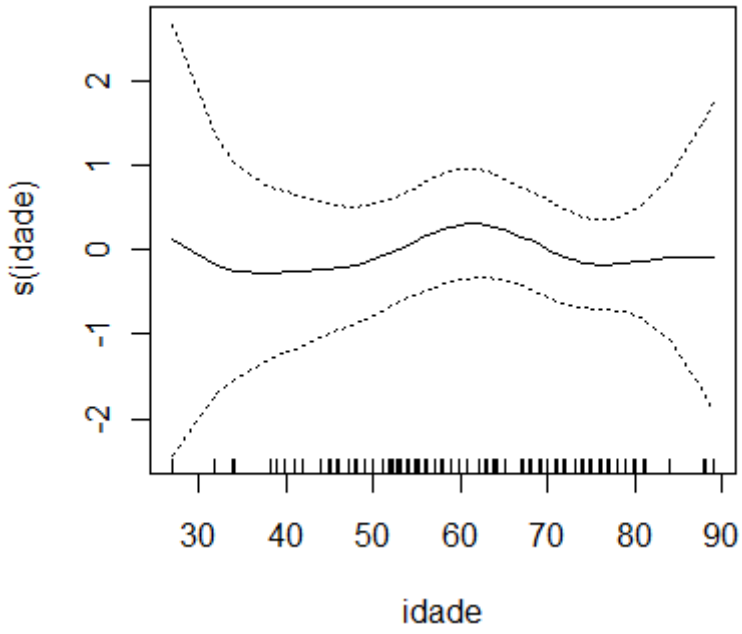

Figura 1: Gráfico de *idade* vs s(*idade*) num modelo linear generalizado aditivo. As curvas a ponteado representam $\pm 2 \times$ desvio padrão pontual.

Passo 4: Introdução das interacções.

Depois de obtido um modelo múltiplo em que todas as variáveis são significativas, para um valor p igual a 0.25, e de verificada a linearidade da covariável idade na escala logit, o passo seguinte é a introdução de interacções.

Tabela 4 Valor obtido para a interacção introduzida.

Covariável	Teste de Wald	Valor p teste de Wald
Idade*escala dor	5.0	0.66

O valor p obtido para a interacção entre a covariável idade e escala da dor não é significativo, tendo em conta os valores habituais de significância, $p<0.05$.

Os dois modelos, com e sem interacção, foram comparados pelo teste de razão de verosimilhanças:

```
Model 1: tempototal ~ idade + escalador + transferido + tipocompanhia +
nívelsócioeconómico + funções + zona
Model 2: tempototal ~ idade * escalador + tipocompanhia +
nívelsócioeconómico + transferido + funções + zona
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       108     120.39
2       101     114.11  7   6.2887    0.5065
```

Tendo em atenção que o valor p é de 0.5065 o modelo sem interacção não é rejeitado. Conclui-se assim que a interacção não é relevante para este conjunto de dados.

Uma vez que a interacção não é significativa tem-se então que o modelo final é dado por:

Modelo Final: Tempo total ~ idade + transferido + zona + funções + nível socioeconómico + tipo de companhia + escala da dor.

Método de selecção: Método s*tepwise*

Começou-se inicialmente pela criação de um modelo logístico com todas as covariáveis incluindo a interacção entre a covariável idade e a escala da dor.

Modelo inicial:

Tempo total ~idade*escalador + sexo + dia + transferido + zona + antecedentes + factores risco + nível socioeconómico+ nível escolaridade + tipo companhia + funções + conhecimento doença.

Modelos obtidos:

Both stepwise:

tempo total ~ escala da dor + zona + nível socioeconómico + funções

Backward stepwise:

tempo total ~ escala da dor + zona + nível socioeconómico + funções

Forward stepwise:

tempo total ~ tipo de companhia + nível socioeconómico + escala da dor + funções

Comparação dos modelos:

```
extractAIC(tempototal.stepback)
15.0000 159.6456
extractAIC(tempototal.stepboth)
15.0000 159.6456
extractAIC(tempototal.stepfor)
17.0000 160.0035
```

Tendo em atenção o critério AIC conclui-se que o modelo obtido pelo método de ***both stepwise*** é o seleccionado pois é o que apresenta menor valor AIC.
Uma vez que a covariável idade não é selecionada pelos métodos de *stepwise*, não é necessário a verificação da linearidade.

Comparação dos modelos obtidos:

Após a aplicação dos diferentes métodos de selecção de covariáveis, "Hosmer e Lemeshow", e ainda pelos métodos de selecção *both, backward* e *forward stepwise,* obteve-se finalmente dois modelos candidatos a modelos explicativos do tempo total de isquémia, um obtido pelo método *both stepwise* e outro pelo método de "Hosmer e Lemeshow.

Como estes dois modelos estão aninhados foram comparados com recurso ao Teste de Razão de Verosimilhanças.

```
Model 1: tempototal ~ escalador + zona + nívelsócioeconómico + funções
Model 2: tempototal ~ idade + escalador + transferido + zona +
nívelsócioeconómico + tipocompanhia + funções
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       113     129.65
2       108     120.39  5   9.2517   0.09944 .
```

A diferença entre os dois modelos é estatisticamente significativa ao nível de significância 5% (valor p=0.09944) pelo que o modelo seleccionado é o obtido pelo método *both stepwise.*

Conclusão:

Quando aplicados os métodos de selecção de s*tepwise* a covariável idade nunca é seleccionada. No caso do modelo seleccionado com base no método de "Hosmer e Lemeshow" a covariável idade aparece embora apresente valores p não significativos para os valores habituais de significância, quer com base no teste de Wald (valor p: 0.15) quer com base no teste de Razão de Verosimilhança (valor p: 0.1486). Este facto não impediu que fosse verificada a linearidade da covariável idade na escala logit e posteriormente feito o ajustamento de um novo modelo com a interacção da idade com a escala da dor, tendo-se verificado que esta interacção não era estatisticamente significativa. Tendo em conta o que se acaba de expor optou-se neste estudo pela escolha do modelo com a covariável idade categorizada, sendo esta última opção a mais reportada na literatura, uma vez que possibilita uma mais fácil comparação entre os dados obtidos neste estudo e os publicados na literatura.

Apêndice 4

Tabela 1: OR e Valor p para a alteração na classe da referência para a covariável idade.

	OR(IC)	Valor p (Wald)
Idade: classe de referência=1		
2	1.37	0.642

Tabela 2: OR e Valor p para as diferentes alterações na classe da referência para a covariável escala da dor.

	OR(IC)	Valor p (Wald)
Escala da dor: classe de referência=4		
5	0.45 (0.06,3.61)	0.453
6	0.61 (0.14,2.65)	0.51
7	0.28 (0.05,1.64)	0.159
8	0.16 (0.03,0.81)	0.027
9	0.12 (0.01,1.09)	0.06
10	0.08 (0.01,0.58)	0.012
Escala da dor: classe de referência=5		
6	1.35 (0.19,9.72)	0.763
7	0.62 (0.07,5.6)	0.67
8	0.36 (0.04,3.05)	0.351
9	0.26 (0.02,3.41)	0.302
10	0.19 (0.02,2.14)	0.177
Escala da dor: classe de referência=6		
7	0.46 (0.09,2.23)	0.334
8	0.27 (0.06,1.18)	0.081
9	0.19 (0.02,1.58)	0.124
10	0.14 (0.02,0.79)	0.026
Escala da dor: classe de referência=7		
8	0.59 (0.1,3.52)	0.559
9	0.41 (0.04,4.3)	0.459
10	0.3 (0.04,2.26)	0.243
Escala da dor: classe de referência=8		
9	0.7 (0.07,6.82)	0.762
10	0.51 (0.08,3.3)	0.481
Escala da dor: classe de referência=9		
10	0.73 (0.06,8.26)	0.797

Tabela 3: OR e Valor p para a alteração na classe da referência para a covariável nível socioeconómico.

	OR(IC)	Valor p (Wald)
Nível Socioeconómico: classe de referência=4		
3	77%	0.01639
2	80%	0.01059
Nível Socioeconómico: classe de referência=3		
2	14.28%,	0.8217

Tabela 4: OR e Valor p para as alterações na classe da referência para a covariável funções.

	OR(IC)	Valor p (Wald)
Funções classe de referência=1		
2	2.88	0.21619
3	1.98	0.3449
Funções: classe de referência=2	5.7	0.0086.
3		

Anexo 1

Escala de Graffar

Classe I: Famílias cuja soma de pontos vai de 5 a 9.

Classe II: Famílias cuja soma de pontos vai de 10 a 13.

Profissão	Grau	Instrução	Grau	Origem do rendimento familiar	Grau	Tipo de habitação	Grau
Grandes empresários; Gestores de topo do sector público e privado (> de 500 empregados); Professores universitários; General, marechal; Profissões liberais (curso superior); dirigentes políticos.	1	Licenciatura; Mestrado; Doutoramento.	1	Lucros de empresas, de propriedades; Heranças.	1	Casa ou andar luxuoso, espaçoso e com máximo de conforto.	1
Médios empresários; Dirigentes de empresas (? de 500 empregados); Agricultores e proprietários; Dirigentes intermédios e quadros técnicos do sector público ou privado; Oficiais das forças armadas; Professores do ensino básico e secundário.	2	Bacharelato.	2	Altos vencimentos e honorários (>=10 vezes o salário mínimo nacional);	2	Casa e andar bastante espaçoso e confortável.	2
Pequenos empresários (? de 50 empregados); Quadros médios; Médios agricultores; Sargentos e equiparados.	3	12º ano; Nove ou mais anos de escolaridade.	3	Vencimentos certos	3	Casa ou andar modesto em bom estado de conservação	3
Pequenos agricultores e rendeiros; Técnicos Administrativos; Operários semi-qualificados; Funcionários públicos e membros das forças armadas ou militarizadas.	4	Escolaridade > 4 anos e < 9 anos	4	Remunerações <= ao salário mínimo nacional; Pensionistas ou reformados; Vencimentos incertos.	4	Casa ou andar degradado.	4
Assalariados agrícolas; Trabalhadores indiferenciados e profissões não classificadas nos grupos anteriores.	5	Escolaridade < 4 anos; Analfabetos.	5	Assistência (subsídios).	5	Impróprio (barraca, andar ou outro); Coabitação de várias famílias em situação de promiscuidade.	5

Classe III: Famílias cuja soma de pontos vai de 14 a 17.

Classe IV: Famílias cuja soma de pontos vai de 18 a 21.

Classe V: Famílias cuja soma de pontos vai de 22 a 25.

Anexo 2

Escala numérica da dor

Sem Dor **Dor Máxima**

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

A Escala Numérica consiste numa régua dividida em onze partes iguais, numeradas sucessivamente de 0 a 10.

Esta régua pode apresentar-se ao doente na horizontal ou na vertical.

Pretende-se que o doente faça a equivalência entre a intensidade da sua Dor e uma classificação numérica, sendo que a 0 corresponde a classificação “Sem Dor” e a 10 a classificação “Dor Máxima” (Dor de intensidade máxima imaginável).